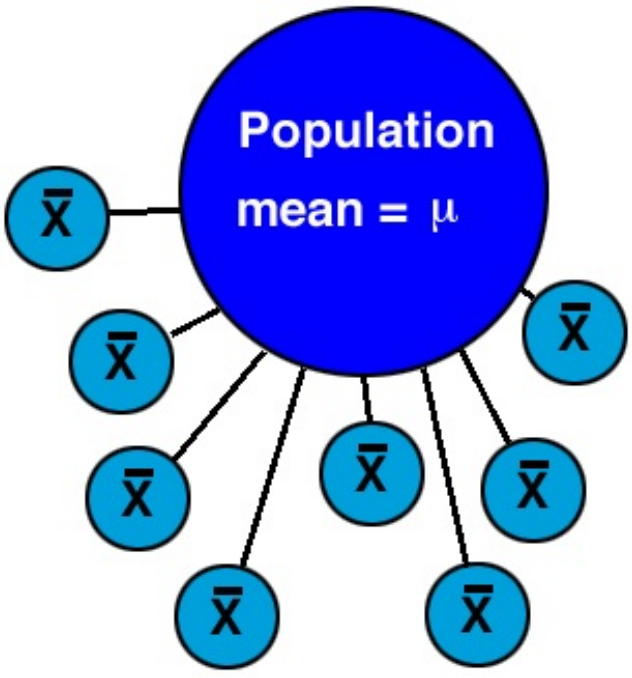
STAT 350 Lecture 5: Sampling Distributions and Central Limit Theorem

Foundations of Statistical Inference

(Chapter 8 of WMMY)

Outline



1. Probability model for the population: $X \sim F$ with unknown mean parameter μ;
2. Get an IID sample of size n: $\{x_1, \ldots, x_n\}$;
3. Use sample average $\bar{x}_n$ to estimate μ;
4. Understand the variability (distribution) of the random variable $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Law of Large Numbers

In general, suppose a sequence of I.I.D. RVs $X_1, X_2, \ldots$ are observed, with common mean μ and common variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the mean of the first n observations $X_1, \ldots, X_n$. Then $E(\overline{X}) = \mu$, $V(\overline{X}) = \sigma^2/n$, and hence $\overline{X}_n$ will be arbitrary close to μ as long as n large enough.

Outline

Sampling from populations

- In statistics, the model is usually defined via a RV X for a population. For example, for all IUPUI students as a population of interest, we may consider a random variable X as the height of an IUPUI student.
- If many individuals are drawn from the population, then we could draw a histogram of the corresponding values for the random variable X of interest to approximate the distribution of X.
- Now suppose we take a sample $X_1, \ldots, X_n$ from this population and calculate some statistic (summary information for the sample), call it T_n, where n indicates the sample size.
- We can again imagine taking many samples of size n from this population, for each sample calculating T_n, and drawing a histogram of these values.
- This histogram is approximating the *sampling distribution* of the statistic T_n.

Motivation Example 1

- Suppose $X_1, X_2, \cdots, X_n$ I.I.D. $X \sim \text{Bin}(1, p)$, which is also called Bernoulli Distribution Ber(p).
- $E(X_i) = p$, $V(X_i) = p(1-p)$.
- With $n = 2$, what is the distribution of $T_n = X_1 + X_2 + \cdots + X_n$? How about $n = 3, 5, 10, 100$?

Motivation Example 2

- Suppose $X_1, X_2, \cdots, X_n$ I.I.D. follow **normal distribution** with mean μ and standard deviation σ.
- With $n = 2$, what is the distribution of $\overline{X}_2$?
- How about $n = 3, 5, 10, 100$?

Motivation Example 3

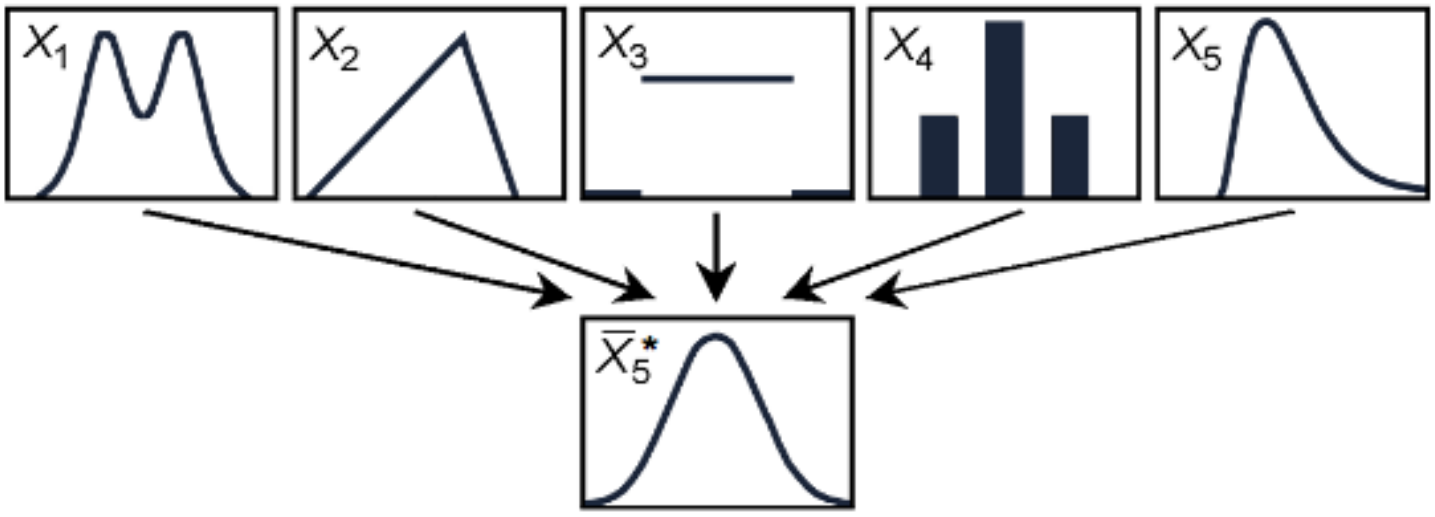
- Suppose $X_1, X_2, \cdots, X_n$ I.I.D. follow the following distribution

x	0	1	2	3
$f(x)$	0.1	0.2	0.4	0.3

- $E(X_i) = 1.9, V(X_i) = 0.89$.
- With $n = 2$, what is the distribution of $\overline{X}_2$? How about $n = 3, 5, 10, 100$?

Outline

Central Limit Theorem



- The fundamental idea is this:

  the average of a large number of RVs with the common mean μ and standard deviation σ has an approximate Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$

CLT

Theorem (Central Limit Theorem)

If $X_1, \ldots, X_n$ are I.I.D. RVs with mean μ and variance $\sigma^2 < \infty$, then for large n both the sample total T_n and the sample mean $\overline{X}_n$ are approximately Normal. In particular, for large n,

$$T_n \sim \mathrm{N}(n\mu, n\sigma^2) \quad \textit{and} \quad \overline{X}_n \sim \mathrm{N}(\mu, \sigma^2/n).$$

- It doesn't matter what kind of RVs $X_1, \ldots, X_n$ are!
- In some sense, sums of independent RVs with a finite variance are "attracted" to Normal.
- How large is "large"? A **rule of thumb** is that the CLT is valid when $n \geq 30$.

Revisit 3 Motivation Examples

- $X_1, X_2, \cdots, X_n$ I.I.D. $\sim$ Bin$(1, p)$:

$$T_n = X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, p) \overset{\text{approx. if } n \geq 30}{\sim} N(np, np(1-p)),$$

$$\overline{X}_n = \hat{p} \overset{\text{approx. if } n \geq 30}{\sim} N(p, \frac{p(1-p)}{n}).$$

- $X_1, X_2, \cdots, X_n$ I.I.D. $\sim$ N(μ, σ^2):

$$T_n = X_1 + X_2 + \cdots + X_n \overset{\text{for any } n}{\sim} N(n\mu, n\sigma^2),$$

$$\overline{X}_n \overset{\text{for any } n}{\sim} N(\mu, \frac{\sigma^2}{n}).$$

- $X_1, X_2, \cdots, X_n$ I.I.D. with common $(E(X_i) = \mu, V(X_i) = \sigma^2)$:

$$T_n = X_1 + X_2 + \cdots + X_n \overset{\text{approx. if } n \geq 30}{\sim} N(n\mu, n\sigma^2),$$

$$\overline{X}_n \overset{\text{approx. if } n \geq 30}{\sim} N(\mu, \frac{\sigma^2}{n}).$$

Exercise 1

Radio signals transmitted to Earth from deep space probes are very weak against the noisy background. Suppose there is a 5% chance that a transmitted bit will be misread. If 300 bits are transmitted, what is the probability that no more than 20 of the bits will be misread?

1. Calculate the exact probability.
2. Use the CLT to approximate the probability.

Exercise 2

Organizers of a fishing tournament believe that the lake holds a sizable population of largemouth bass. They assume that the weight of largemouth bass follows a right-skewed model with mean 3.5 lbs and Std. Dev 2.5 lbs.

1. Do we have enough information to calculate the probability that a randomly caught largemouth bass weighs under 4 lbs?
2. Each fisherman catches 49 fish, what is the probability a single fisherman's catch will average under 4 lbs?
3. In order to be in top 10% in this tournament, at least how many lbs the fisherman should have on average for the 49 fish?

Exercise 3

Roll a fair six-sided die 50 times and let $\overline{X}$ denote the average of the rolls.

1. What is the probability that the average is between 3 and 4?
2. How many times should you roll the die to be 95% certain the average will fall between 3 and 4?

CLT for two sample case

Theorem (Central Limit Theorem (two sample))

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

- If $n_1 \geq 30$ and $n_2 \geq 30$, the normal approximation is very good when the underlying distributions are not too far away from normal.
- If both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution no matter what the sizes of n_1 and n_2 are.

Example

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the drying time for the two types of paint are both normally distributed with the same mean drying time, find

$$\mathrm{P}(\bar{X}_A - \bar{X}_B > 1).$$

Example

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.
Assuming that the drying time for the two types of paint are both normally distributed with the same mean drying time, find

$$P(\bar{X}_A - \bar{X}_B > 1).$$

- The machinery in the calculation is based on the presumption that $\mu_A = \mu_B$. Suppose, however, that the experiment is actually conducted for the purpose of drawing an inference regarding the equality of μ_A and μ_B, the two population mean drying times.
- If the two averages differ by as much as 1 hour (or more), this clearly is evidence that would lead one to conclude that the population mean drying time is not equal for the two types of paint.

Outline

Setup

- It's a consequence of the CLT that n big enough

$$\overline{X} \sim \mathsf{N}(\mathsf{E}(\overline{X}) = \mu, \mathsf{V}(\overline{X}) = \sigma^2/n),$$

  hence the distribution of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

  is the standard Normal, $\mathsf{N}(0, 1)$.
- But it is often the case that the population variance σ^2 is *unknown*.
- A natural idea is to replace it with the sample variance S^2. And if sample size n is big, we indeed can do this.

Result

- However, we know that if $X_1, \ldots, X_n$ are Normal

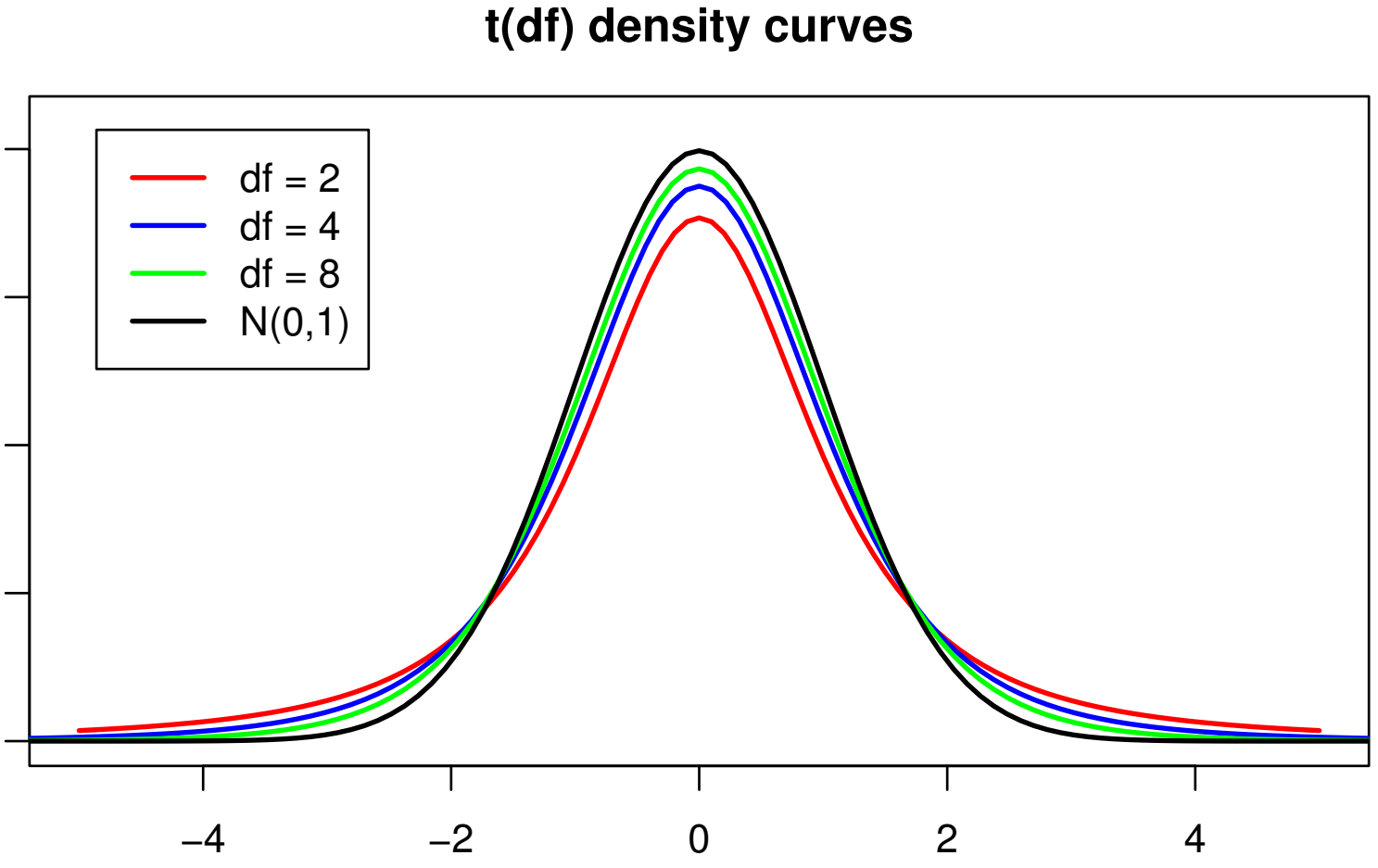
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

  is always Normal no matter how big of the sample size n!
- How about the following with n small

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

- Due to the randomness brought by the sample variance S^2, such a random variable is not normal any more, but it has a *Student-t* distribution, written as $T \sim \mathrm{t}_{n-1}$, where the subscript denotes the *degrees of freedom* parameter.

PDFs of Student-t distribution



Exercise

Suppose $T \sim \mathrm{t}_{10}$.

1. Find the 90th percentile.
2. Find the 77th percentile.
3. Find $\mathrm{P}(T \geq 1.75)$.
4. Find $\mathrm{P}(1 < T \leq 2)$.
5. Find $\mathrm{P}(T < -2.112)$.

Summary—Most Important to Keep in Mind

1. without normality, $n \geq 30$ and σ is given: $\frac{(\overline{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$
2. without normality, $n \geq 30$ and s is given: $\frac{(\overline{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{N}(0,1)$
3. X_i is Normal, σ is given: $\frac{(\overline{X}_n - \mu)}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$
4. X_i is Normal, s is given: $\frac{(\overline{X}_n - \mu)}{s/\sqrt{n}} \sim \mathrm{t}(n-1)$
5. X_i is Bernoulli, $n \geq 30$: $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim \mathrm{N}(0,1)$

Note that Bernoulli random variable is just the Binomial random variable with only 1 trial.