AFRE 802
Statistical Methods for Agricultural, Food, & Resource Economists

Sampling distributions & the Central Limit Theorem – Part 1 of 2

(WMS Ch. 7.1-7.2)

October 24, 2017

Nicole Mason
Michigan State University
Fall 2017

GAME PLAN

Return and briefly discuss midterm

Overview of topics in the rest of the course

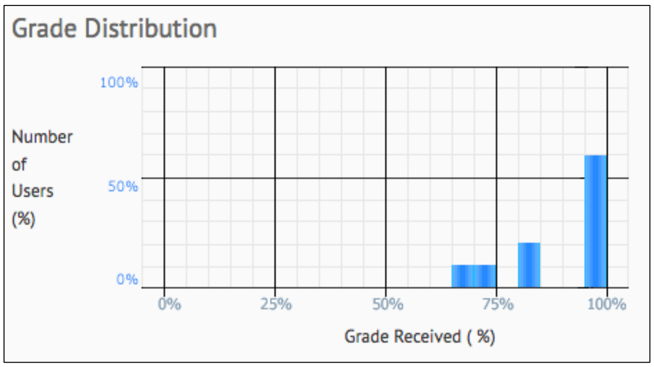
Start discussion of sampling distributions (Part 1 of 2)

1. Introduction
2. Sampling distributions related to the normal distribution
   a. Review of standard normal RVs & link to sample mean
   b. Review of chi-square RVs & link to sample variance
   c. The *t* distribution (also related to chi-square RVs)
   d. The *F* distribution (also related to chi-square RVs)

1

Midterm results

- Grades distribution (in % terms)
  - Mean = 88%
  - Median = 95%
  - Clusters of grades
    66-71, 82-83, 95+
- Expect a somewhat more challenging final exam ☺

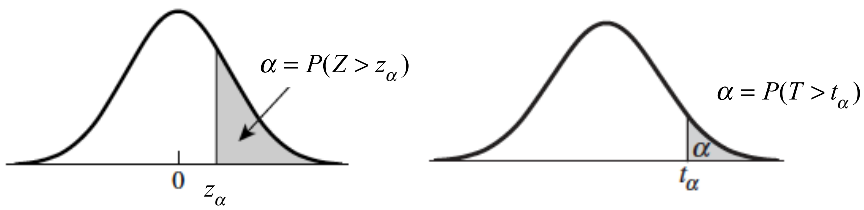


2

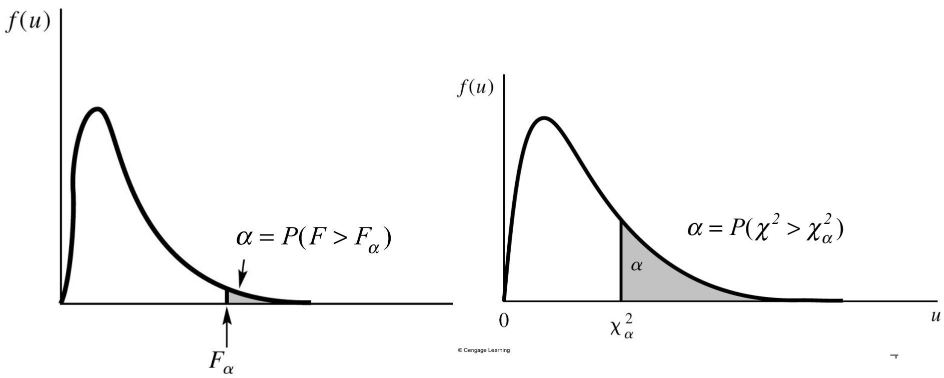
What we'll be covering during the second half of the course:

- Statistics and their distributions (sampling distribution = probability distribution of a statistic)
- Two important theorems:
  - The Law of Large Numbers
  - The Central Limit Theorem
- Estimating population parameters based on a sample of data
- Testing hypotheses about those population parameters
- Intro to linear regression

3



Sampling distributions: chi-squared, t, and F



Motivation

MICHIGAN STATE UNIVERSITY

- Recall that a major **objective of statistics** is to **make an inference about a population based on info in a sample** from that population
  - EX) Use sample mean to draw inferences about population mean
- If we draw a **random sample of *N*** observations from a given population of interest, then **each observation, y_1, y_2, …, y_N,** is the **realization of** its corresponding **random variable Y_1, Y_2, …, Y_N**
- These RVs are independent and have the same distribution ("**independent & identically distributed", i.i.d.**)
- Before the midterm, we studied linear functions of RVs. The sample mean is a linear function of RVs:

$$\boxed{\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i}$$

- Therefore, it too is an RV.
- It is also a **statistic** (a function of observable RVs in a sample & known constants). The probability distribution of a statistic is called a **sampling distribution**.

MICHIGAN STATE UNIVERSITY

Motivation (cont'd)

- *Other than the sample mean, what are some other examples of statistics?*
  - Sample variance
  - Sample median
  - Sample range
  - Sample minimum
  - Sample maximum
  - Sample 90^{th} (or whatever) percentile
  - Etc.

The sampling distribution of the sample mean

If $Y_1, Y_2, \ldots, Y_N$ is a random sample of size N **from a normal distribution** with mean, μ, and variance, σ^2, then the **sample mean**

$$\boxed{\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i}$$

is **normally distributed** with

$$\boxed{\text{mean } \mu_{\bar{Y}} = \mu \quad \text{and} \quad \text{variance } \sigma^2_{\bar{Y}} = \frac{\sigma^2}{N}}$$

YouTube videos with simulations:

- Start at 1:38 and watch until 8:00

https://www.youtube.com/watch?v=BwE2a18Th4c&feature=youtu.be

- Sampling distribution of the sample mean: watch through 4:42 for case where original RV is normally distributed (***watch on your own after class***): https://www.youtube.com/watch?v=sZ0DsE4vhqk

The sampling distribution of the sample mean

If $Y_1, Y_2, \ldots, Y_N$ is a random sample of size N **from a normal distribution** with mean, μ, and variance, σ^2, then the **sample mean**,

$$\boxed{\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i}$$

is **normally distributed** with

$$\boxed{\text{mean } \mu_{\bar{Y}} = \mu \quad \text{and} \quad \text{variance } \sigma_{\bar{Y}}^2 = \frac{\sigma^2}{N}}$$

I suggest you work through these proofs for practice.

How could we convert this to a standard normal RV, Z?

$$\boxed{Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / N}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{N}}, \quad \text{where } Z \sim N(0,1)}$$

MICHIGAN STATE UNIVERSITY

$$\boxed{Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{N}}, \; Z \sim N(0,1)}$$

EXAMPLE 7.2

A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It has been observed that the amount of fill dispensed by the machine is normally distributed with $\sigma = 1.0$ ounce. A sample of $n = 9$ filled bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting), and the ounces of fill are measured for each. Find the probability that the sample mean will be within .3 ounce of the true mean μ for the chosen machine setting.

9

The sampling distribution of the sum of squares of N independent standard normal RVs: χ^2

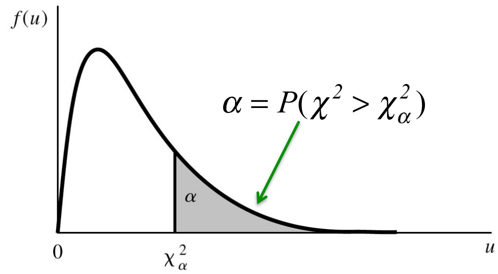
Let $Z_i = \dfrac{Y_i - \mu}{\sigma}$, $i = 1, 2, \ldots, N$ be independent standard normal RVs. Then

$$\sum_{i=1}^{N} Z_i^2 = \sum_{i=1}^{N} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

has a χ^2 distribution with N degrees of freedom (d.f.)

See p. 356 and p. 321 in WMS for proof

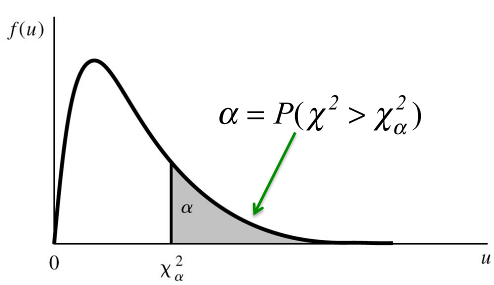
See Table 6 in Appendix 3 for table of probabilities for the chi-square distribution.



EXAMPLE 7.4

If $Z_1, Z_2, \ldots, Z_6$ denotes a random sample from the standard normal distribution, find a number b such that

$$P\left(\sum_{i=1}^{6} Z_i^2 \le b\right) = .95.$$



11

The sample variance & link to the chi-square dist.

- Recall the formula for the sample variance for a random sample of size *N*:

$$\boxed{S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

- Move (*N-1*) to the left side and divide both sides by σ^2:

$$\frac{(N-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

has a χ^2 distribution with (N - 1) d.f.

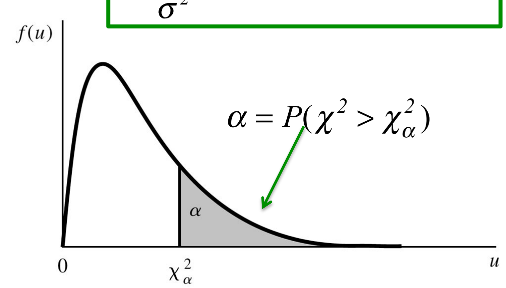
See p. 358 in WMS for proof

EXAMPLE **7.5**

In Example 7.2, the ounces of fill from the bottling machine are assumed to have a normal distribution with $\sigma^2 = 1$. Suppose that we plan to select a random sample of ten bottles and measure the amount of fill in each bottle. If these ten observations are used to calculate S^2, it might be useful to specify an interval of values that will include S^2 with a high probability. Find numbers b_1 and b_2 such that

$$P(b_1 \le S^2 \le b_2) = .90.$$

$$\boxed{\frac{(N-1)S^2}{\sigma^2} \sim \chi^2 \text{ with } (N-1) \text{ d.f.}}$$



(Student's) *t* distribution & replacing *σ* with *S* in our standardized statistic for the sample mean

- Earlier today we looked at:

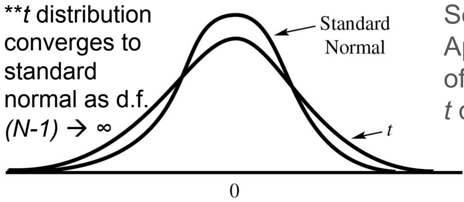
$$\boxed{Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{N}} \quad \sim N(0,1)}$$

- If we don't know *σ* (which we often don't), we can replace it with *S* to get a new statistic, *T*:

$$\boxed{T = \frac{\overline{Y} - \mu}{S / \sqrt{N}} \quad \sim t \text{ distribution with } N - 1 \text{ d.f.}}$$

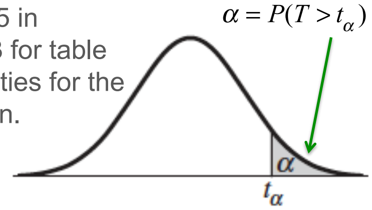
We'll use this later to test hypotheses related to *μ*.

**t distribution converges to standard normal as d.f. *(N-1)* → ∞



See Table 5 in Appendix 3 for table of probabilities for the *t* distribution.

$\alpha = P(T > t_\alpha)$



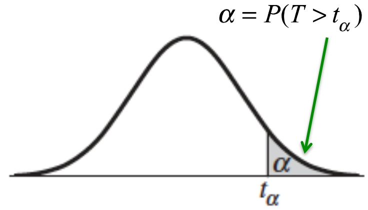
MICHIGAN STATE UNIVERSITY

EXAMPLE 7.6

The tensile strength for a type of wire is normally distributed with unknown mean μ and unknown variance σ^2. Six pieces of wire were randomly selected from a large roll; Y_i, the tensile strength for portion i, is measured for $i = 1, 2, \ldots, 6$. The population mean μ and variance σ^2 can be estimated by $\overline{Y}$ and S^2, respectively. Because $\sigma^2_{\overline{Y}} = \sigma^2/n$, it follows that $\sigma^2_{\overline{Y}}$ can be estimated by S^2/n. Find the approximate probability that $\overline{Y}$ will be within $2S/\sqrt{n}$ of the true population mean μ.

$$\boxed{T = \frac{\overline{Y} - \mu}{S / \sqrt{N}} \quad \sim t \text{ with } (N - 1) \text{ d.f.}}$$

$\alpha = P(T > t_\alpha)$



MICHIGAN STATE UNIVERSITY

t distributions in general

Let Z be a standard normal random variable and let W be a χ^2-distributed variable with ν df. Then, if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

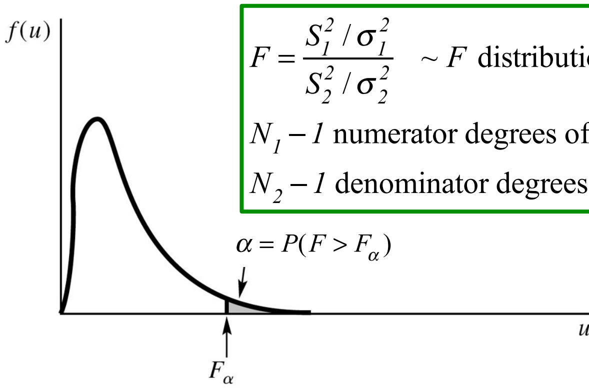
is said to have a *t distribution* with ν df.

16

Comparing the variances of 2 normal populations and the *F* distribution

- When testing hypotheses about the means of 2 (potentially different) normal populations, we often need to compare the variances of those two populations.
- We use *F* statistics to do this (assuming we have independent random samples from the two populations):

$$\boxed{\begin{array}{l} F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \quad \sim F \text{ distribution with} \\ N_1 - 1 \text{ numerator degrees of freedom and} \\ N_2 - 1 \text{ denominator degrees of freedom} \end{array}}$$



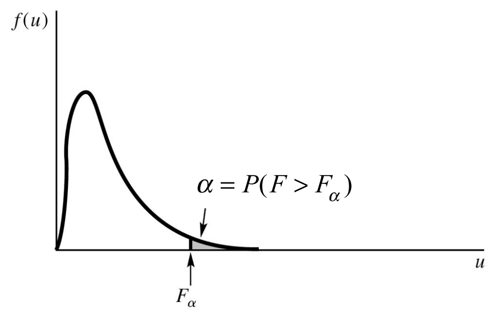
See Table 7 in Appendix 3 for table of probabilities for the *F* distribution.

17

EXAMPLE 7.7

If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal population variances, find the number b such that

$$P\left(\frac{S_1^2}{S_2^2} \leq b\right) = .95.$$



$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F \text{ with } (N_1 - 1) \text{ numerator d.f. \& } (N_2 - 1) \text{ denominator d.f.}$$

MICHIGAN STATE UNIVERSITY

F distributions in general

Let W_1 and W_2 be *independent* χ^2-distributed random variables with ν_1 and ν_2 df, respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

is said to have an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom.

19

Summary

If $Y_1, Y_2, \ldots, Y_N$ is a **random sample** of size N **from a normal distribution** with mean, μ, and variance, σ^2, then:

$$\boxed{Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{N}} \sim N(0,1)}$$

$$\boxed{\frac{(N-1)S^2}{\sigma^2} \sim \chi^2 \text{ with } (N-1) \text{ d.f.}}$$

$$\boxed{T = \frac{\bar{Y} - \mu}{S / \sqrt{N}} \sim t \text{ with } (N-1) \text{ d.f.}}$$

If we have two independent random samples from normal populations with variances σ_1^2 and σ_2^2, then:

$$\boxed{F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F \text{ with } (N_1 - 1) \text{ numerator d.f. \& } (N_2 - 1) \text{ denominator d.f.}}$$

20

MICHIGAN STATE UNIVERSITY

Homework:

- WMS Ch. 7 (part 1 of 2)
  - 7.9, 7.10 (part a only), 7.11, 7.12, 7.15, 7.19, 7.26, 7.31 (parts a-c only)
  - Not graded but I strongly encourage you to try the extra in-class exercises at the end of the PPT for more practice.

•We'll likely finish Ch. 7 on Thursday, so Ch. 7 HW will be due next Tuesday (Oct. 31)

Next class:

- Sampling distributions & Central Limit Theorem (Part 2 of 2)

Reading for next class:

- WMS Ch. 7 (sections 7.3, 7.5-7.6)

FYI in D2L:

- YouTube videos on using Stata to compute Z, t, F, and chi-square probabilities

21

MICHIGAN STATE UNIVERSITY

In class exercise #1: work through the proof of the mean and variance of the sample mean

22

MICHIGAN STATE UNIVERSITY

In class exercise #2

EXAMPLE **7.3**

Refer to Example 7.2. How many observations should be included in the sample if we wish $\overline{Y}$ to be within .3 ounce of μ with probability .95?

23