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

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

(WMS Ch. 7.3, 7.5-7.6)

October 26, 2017

Nicole Mason
Michigan State University
Fall 2017

GAME PLAN

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

1. Review of the sampling distribution of the sample mean when the underlying RVs ~ normal
2. Review of chi-squared, t, and F distributions
3. The Central Limit Theorem
4. The Law of Large Numbers
5. The normal approximation to the binomial distribution

1

Review

MICHIGAN STATE UNIVERSITY

Sampling distributions related to the sample mean & sample variance of a random sample of normal RVs:

If $Y_1, Y_2, \ldots, Y_N$ is a **random sample from a <u>normal</u> distribution** with mean, μ, and variance, σ^2, then:

- **<u>Sample mean</u>**

$$\boxed{\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \sim Normal(\mu, \frac{\sigma^2}{N})}$$

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

- **<u>Sample variance</u>**

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

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

2

Review: t distribution & replacing σ with S in our standardized statistic for the sample mean

- We just reviewed that **<u>when our random sample is from a normal distribution</u>**, then:

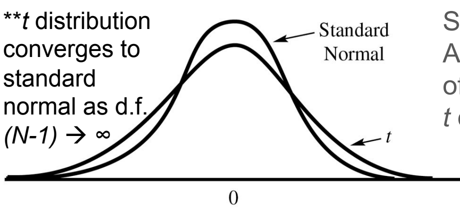
$$\boxed{Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{N}} \sim N(0,1)}$$

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

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

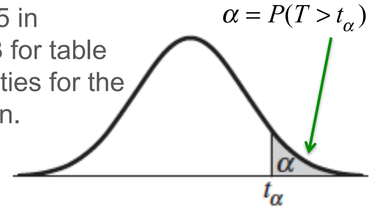
We'll use this later to test hypotheses related to μ.

**t distribution converges to standard normal as d.f. $(N-1) \rightarrow \infty$



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

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



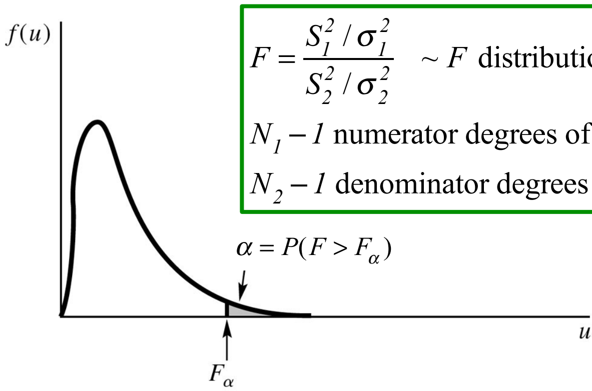
Review: Comparing the variances of 2 normal populations and the *F* distribution

- 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 normal populations):

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F \text{ distribution with}$$

$N_1 - 1$ numerator degrees of freedom and

$N_2 - 1$ denominator degrees of freedom



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

4

Two of the most famous theorems in probability

1. **Central Limit Theorem**

2. **Law of Large Numbers**

5

The Central Limit Theorem (CLT)

- So far we've focused on random samples (Y_i, $i=1,\ldots,N$) drawn from normal distributions
- But we often don't know the distribution from which our random sample is drawn (or our sample if from a non-normal pouplation)
- The **Central Limit Theorem** allows us to approximate the sampling distribution of $\bar{Y}$ regardless of the distribution of Y_i
- **CLT**: The **sampling distribution of the sample mean will be approximately normal as $N \rightarrow \infty$ regardless of the distribution of Y_i**

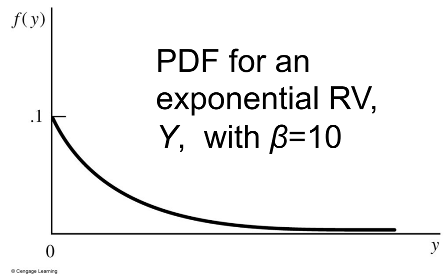
6

CLT – YouTube videos – simulation & review

- https://www.youtube.com/watch?v=BwE2a18Th4c&feature=youtu.be
  - Watch from 8:00 to end
- https://www.youtube.com/watch?v=sZ0DsE4vhqk
  - Watch from 4:44 to end

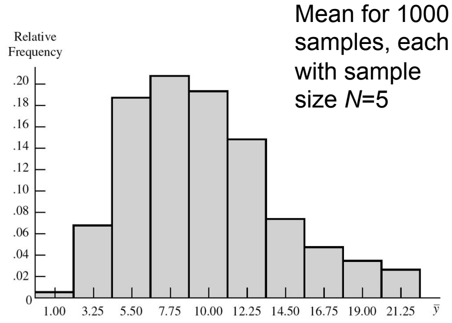
7

Another CLT example before we state it formally

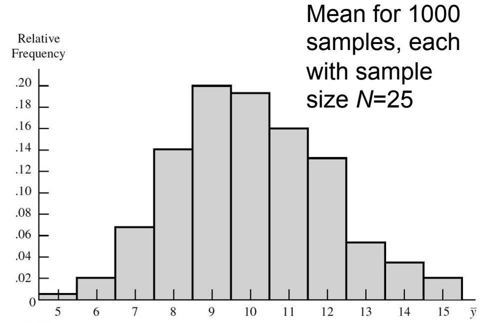


PDF for an exponential RV, Y, with β=10

What is the mean and variance of Y?



Mean for 1000 samples, each with sample size N=5



Mean for 1000 samples, each with sample size N=25

The Central Limit Theorem (CLT) - formally

Let $Y_1, Y_2, \ldots, Y_N$ be i.i.d. distributed RVs with

$E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$, then

the distribution of $\dfrac{\bar{Y}_N - \mu}{\sigma / \sqrt{N}}$

converges to the standard normal as $N \to \infty$

Simpler statement:

$\bar{Y}_N$ is asymptotically normally distributed

with mean μ and variance $\dfrac{\sigma^2}{N}$

See WMS section 7.4 for the proof if interested

- "Large" sample size: roughly ***N>30***
- Note: CLT applies to a random sample from **ANY distribution** with finite mean & variance & large N

9

CLT: example #1

EXAMPLE **7.8**

CLT: Let $Y_1, Y_2, \ldots, Y_N$ be i.i.d. distributed RVs with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$, then the distribution of $\dfrac{\bar{Y}_N - \mu}{\sigma / \sqrt{N}}$ converges to the standard normal as $N \to \infty$

Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of $n = 100$ students from one large high school had a mean score of 58. Is there evidence to suggest that this high school is inferior? (Calculate the probability that the sample mean is at most 58 when $n = 100$.)

μ σ^2

$$P(\bar{Y}_{N=100} \le 58)$$

$$\bar{Y} \approx N\left(\mu = 60, \frac{\sigma^2}{N} = \frac{64}{100}\right)$$

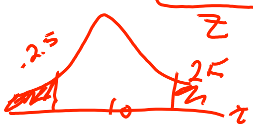
$\mu_{\bar{Y}}$, $\sigma^2_{\bar{Y}}$

$$\approx P\Bigg(\underbrace{\frac{\bar{Y} - \mu}{\sigma/\sqrt{N}}}_{Z} \le \frac{58 - \mu}{\sigma/\sqrt{N}}\Bigg) = P(Z \le -2.5)$$

60, 8, 10

$$= P(Z \ge 2.5)$$

$$= 0.0062$$



CLT: example #2

EXAMPLE **7.9**

CLT: Let $Y_1, Y_2, \ldots, Y_N$ be i.i.d. distributed RVs with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$, then the distribution of $\dfrac{\bar{Y}_N - \mu}{\sigma / \sqrt{N}}$ converges to the standard normal as $N \to \infty$

The service times for customers coming through a checkout counter in a retail store are independent random variables with mean 1.5 minutes and variance 1.0. Approximate the probability that 100 customers can be served in less than 2 hours of total service time.

$$\frac{120 \text{ min}}{100 \text{ people}}$$

$\mu = 1.5$

$\sigma^2 = 1 \Rightarrow \sigma = 1$

$N = 100$

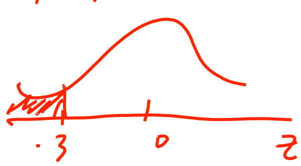
$$P(\bar{Y} < 1.2)$$

$$\approx P(Z < -3)$$

$$= P(Z > 3) = 0.00135$$

$$\parallel$$

$P(Z \ge 3)$ b/c continuous



The Law of Large Numbers

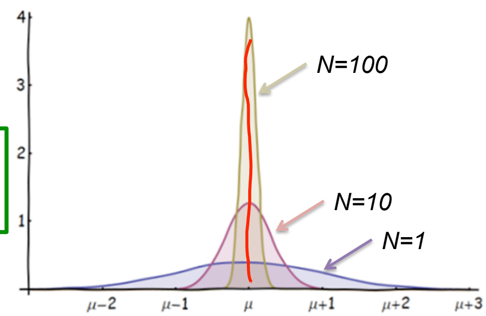
- We often want to know the probability that the sample mean (based on a given sample size, *N*) is within or outside of a certain distance of the population mean:

$$P(|\overline{Y}_N - \mu| < \varepsilon) \quad \text{or} \quad P(|\overline{Y}_N - \mu| > \varepsilon) \quad \text{for some } \varepsilon > 0$$

- *What happens to the latter probability as the sample size (N) gets large?*

$$\boxed{P(|\overline{Y}_N - \mu| > \varepsilon) \to 0 \text{ as } N \to \infty \text{ for any } \varepsilon > 0.}$$

The **"law of large numbers"**: as the sample size goes to infinity, the sample mean converges (in probability) to the population mean.



12

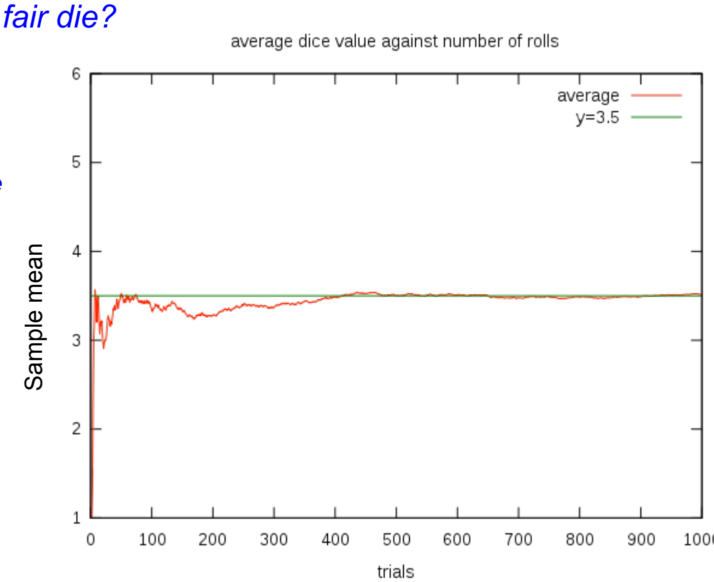
Another illustration of the Law of Large Numbers

- *What is the population mean of the number obtained on a roll of a fair die?*

(1+2+3+4+5+6)/6
=21/6=3.5

Would the sample mean be exactly 3.5 if you rolled the die only a small number of times?

What if you rolled it 1,000 times? A million times?



Normal approximation to the binomial distribution

- Recall that a binomial RV, *Y*, is the # of successes in *n* trials, where the *P(success)* on one trial is *p*

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

- OR think of as *Y* as the sum of *n* binary variables

$$\frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad X_i = \begin{cases} 1, & \text{if the } i\text{th trial results in success,} \\ 0, & \text{otherwise.} \end{cases}$$

- Divide both sides by *n*: $\frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$

- As *n* gets large, by the CLT:

$$\boxed{\frac{Y}{n} = \bar{X} \sim Normal(p, \frac{pq}{n})}$$

σ^2/n

Why is this the mean & variance?

Note: This approximation works well if:

$$n > 9\left(\frac{\text{larger of } p \text{ and } q}{\text{smaller of } p \text{ and } q}\right)$$

MICHIGAN STATE UNIVERSITY

Normal approximation to the binomial

$$\boxed{\frac{Y}{n} = \bar{X} \sim Normal(p, \frac{pq}{n})}$$

This is the distribution of Y/n. What is the distribution of Y?

$Y \sim Normal(np, npq)$

15

$$\frac{Y}{n} = \bar{X} \sim Normal\left(p, \frac{pq}{n}\right)$$

Normal approximation to the binomial – example

EXAMPLE **7.10**

$n > 9\left(\frac{0.5}{0.5}\right)$

Candidate A believes that she can win a city election if she can earn at least 55% of the votes in precinct 1. She also believes that about 50% of the city's voters favor her. If $n = 100$ voters show up to vote at precinct 1, what is the probability that candidate A will receive at least 55% of their votes?

Success = voter votes for candidate A

p = probability of success = 0.5

$n = 100$

$P\left(\frac{Y}{n} = \bar{X} \geq 0.55\right)$

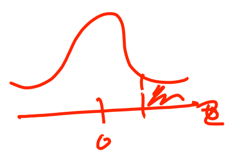
$\bar{X} \approx N\left(0.5, \frac{(0.5)(0.5)}{100}\right)$

$= 0.0025 = \sigma^2$

$\Rightarrow \sigma = 0.05$

$\approx P\left(Z \geq \frac{0.55 - 0.5}{0.05}\right)$

$P(Z \geq 1) = 0.1587$



16

Summary of Chapter 7

- To make inferences about population parameters from info in a random sample from that population, we need to know the probability distributions for the statistics we are using
  - **Statistic** = a function of observable RVs in the sample and known constants
- The probability distributions of statistics are called **sampling distributions**
- We studied several sampling distributions related to the normal distribution, the sample mean, & the sample variance: **standard normal, chi-square, *t*, and *F***
- Also studied the **law of large numbers** and the **central limit theorem** (CLT), and used the CLT to come up with the normal approximation to the binomial
- Up next: estimating population parameters from the info in a sample. After that: hypothesis testing

17

Use remaining class time (if any) to go over extra in-class exercises in previous lecture

Homework:

- WMS Ch. 7 (part 2 of 2)
  - Central limit theorem: 7.43, 7.44, 7.45, 7.48, 7.49
  - Normal approximation to the binomial: 7.71, 7.75
- **Ch. 7 HW is due on Tuesday

Next class

- Estimation (Part 1 of 2) – hurray!

Reading for next class:

- WMS Ch. 8 (sections 8.1-8.4)

MICHIGAN STATE UNIVERSITY

Extra slides

20

Normal approx. to the binomial - continuity correction

- FYI – not something I will test you on
- Binomial is discrete RV, normal is continuous RV
- → If using a continuous RV to approximate a discrete one, often have to make "**continuity correction**"
- EX) If want to know $P(Y \leq 3)$, where $Y \sim$ binomial, use $P(Y \leq 3.5)$ when convert Z . (Recall that binomial RV Y is the # of successes in N trials, so it's an integer.)

See WMS pp. 380-382
and these links for details if you are interested:
http://www.statisticshowto.com/what-is-the-continuity-correction-factor/

http://courses.wcupa.edu/rbove/Berenson/10th%20ed%20CD-ROM%20topics/section6_5.pdf