Math 243

Distribution of Sample Means – Inv. 2.4

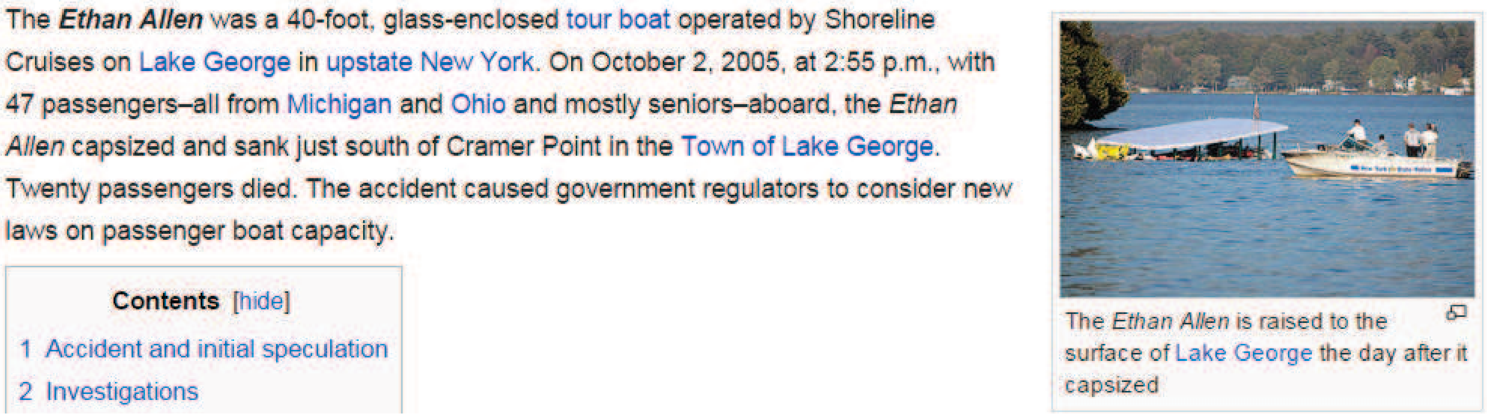
Statistical Inference for 1 Quantitative Variable

Suppose we want to make *inferences* beyond the sample data

- Need random sample from population/process
- Need to how about the *behavior* of sample means from different random samples from the same population

Investigation 2.4 (p. 143)

- Wikipedia



The ***Ethan Allen*** was a 40-foot, glass-enclosed tour boat operated by Shoreline Cruises on Lake George in upstate New York. On October 2, 2005, at 2:55 p.m., with 47 passengers–all from Michigan and Ohio and mostly seniors–aboard, the *Ethan Allen* capsized and sank just south of Cramer Point in the Town of Lake George. Twenty passengers died. The accident caused government regulators to consider new laws on passenger boat capacity.

Contents [hide]

1 Accident and initial speculation
2 Investigations

The *Ethan Allen* is raised to the surface of Lake George the day after it capsized

Inv. 2.4: Try parts (a) – (g) in class

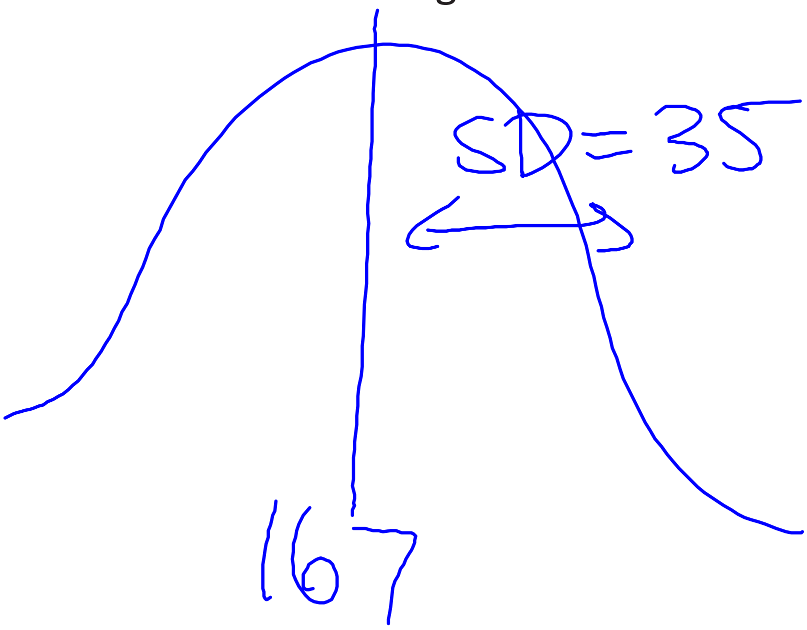
Part (a) and (b): Sketch your idea of the following distributions

random

Weights of 47 adults

similar

Weights of all adults



Inv. 2.4, part (c): if the tour boat company consistently accepted 47 passengers, what is the probability their combined weight exceeds 7500 pounds?

What are the ***observational units***?

47 passengers

What is the ***variable***? What is the ***type*** of the variable?

total weight

Quantitative

Inv. 2.4, part (d). Translate the question from the *total* weight of 47 passengers to the *average* weight of 47 passengers

$$\frac{\text{Total weight of 47 passengers}}{47} > \frac{\text{7500 lbs}}{47}$$

Average weight of 47 passengers > 15957 lbs.

So, to see how often the boat was sent out with too much weight, we need to know about the distribution of the *average weight of 47 passengers* from different boats (samples). Think about a distribution of sample mean weights from different random samples of 47 passengers, selected from the population of adult Americans.

(e) Where do you think the distribution of sample means will be centered? 167

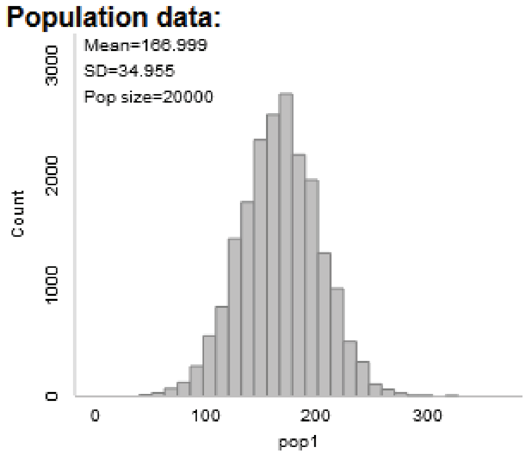
(f) Do you think the distribution of sample means would have more variability, [✓ less variability], or the same variability as weights of individual people? Explain your answer.

(g) Do you think the probability of having the *average* weight exceed 159.574 pounds is larger or smaller than the probability of the weight of any *one* passenger exceeding 159.574 pounds?

Notation for one quantitative variable

μ = population mean

σ = population standard deviation



Notation for one quantitative variable

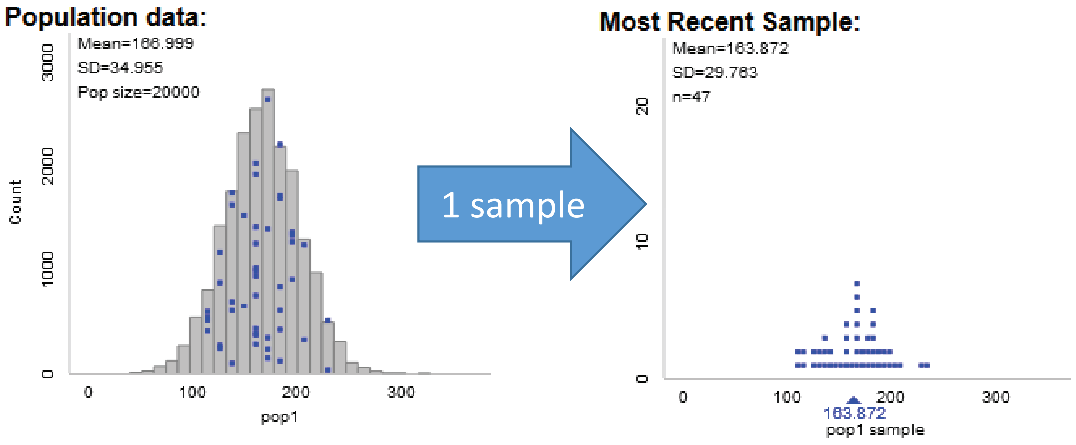
μ = population mean

σ = population standard deviation

$\bar{x}$ = mean of your sample

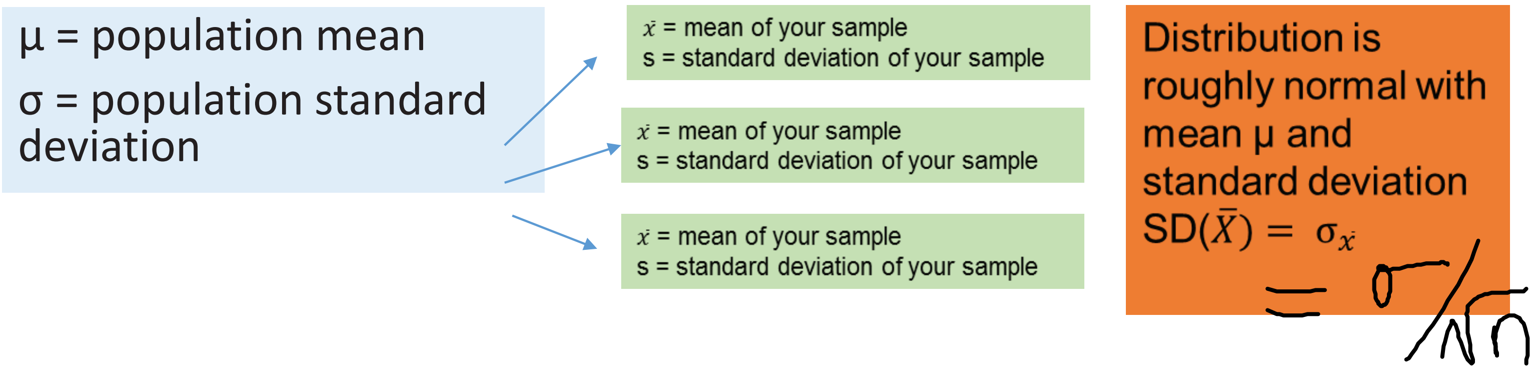
s = standard deviation of your sample

n= sample size



$$\frac{\sigma}{\sqrt{n}} = \frac{35}{\sqrt{47}} = 5,10$$

Notation for one quantitative variable



μ = population mean

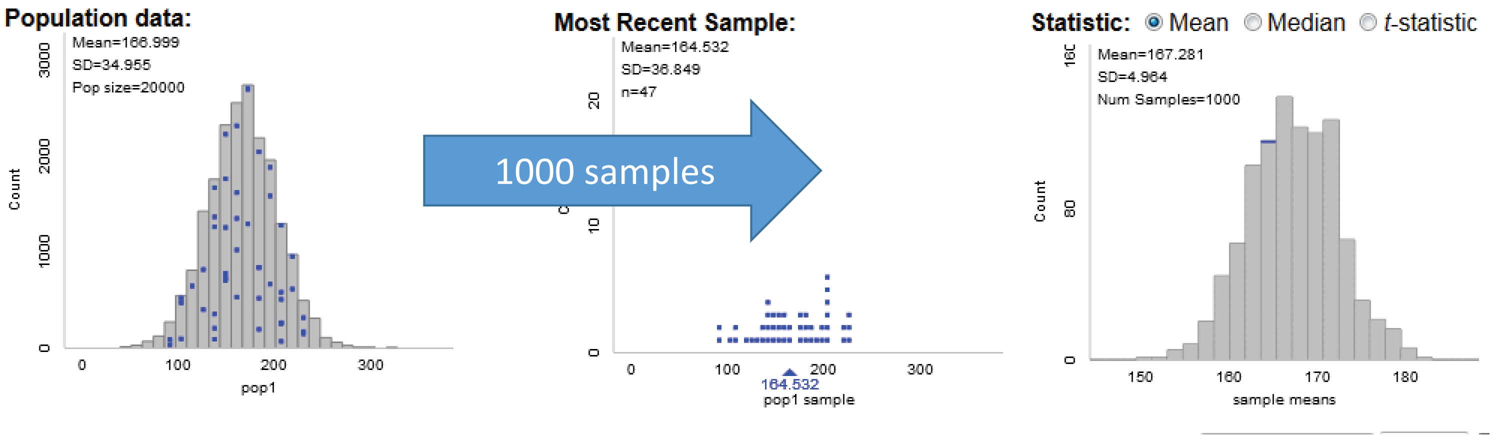
σ = population standard deviation

$\bar{x}$ = mean of your sample
s = standard deviation of your sample

$\bar{x}$ = mean of your sample
s = standard deviation of your sample

$\bar{x}$ = mean of your sample
s = standard deviation of your sample

Distribution is roughly normal with mean μ and standard deviation SD($\bar{X}$) = $\sigma_{\bar{x}}$ $= \sigma/\sqrt{n}$



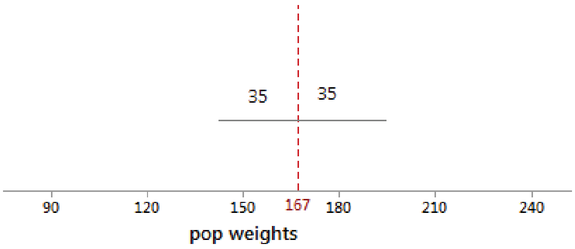
Population data:

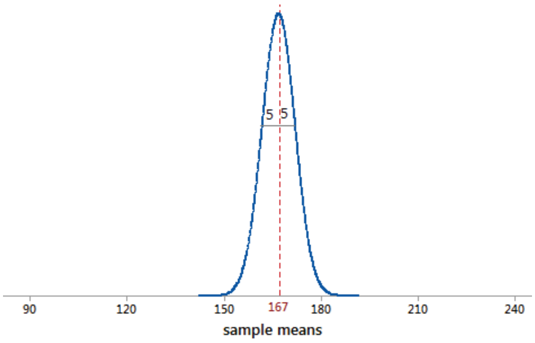
Most Recent Sample:

Statistic: ● Mean ○ Median ○ *t*-statistic

Investigation 2.4: *Ethan Allen*

- CDC: population mean 167 lbs, population standard deviation 35 lbs
- Distribution of **sample means**
  - Normal or Approximately normal assuming the population of weights is not strongly skewed, probably safe assumption with weights of humans
  - With mean 167 lbs and sd = 35/sqrt(47) = 5.11 lbs





Central Limit Theorem for sample means

For large populations or processes with long-run/population mean μ and standard deviation σ, the distribution of sample means has:

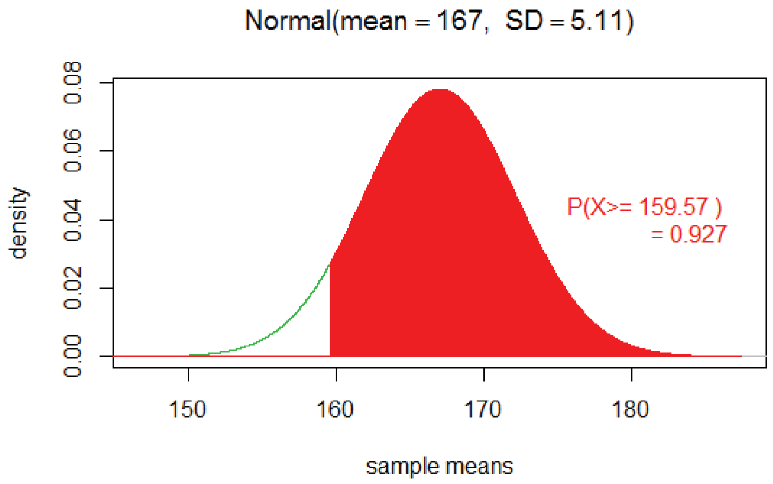
- Mean = $\underline{\mu}$
- Standard deviation = $\underline{\sigma/\sqrt{n}}$
- Shape
  - Normal if population is normal
  - Approximately normal if sample size is large

$n \geq 30$

$$P\left(\frac{\text{total}}{47} > \frac{7500}{47}\right)$$

Investigation 2.4: part n

- So $P(\bar{X} \geq \boxed{159.57}) = .927$



- About 93% of boats were overweight…

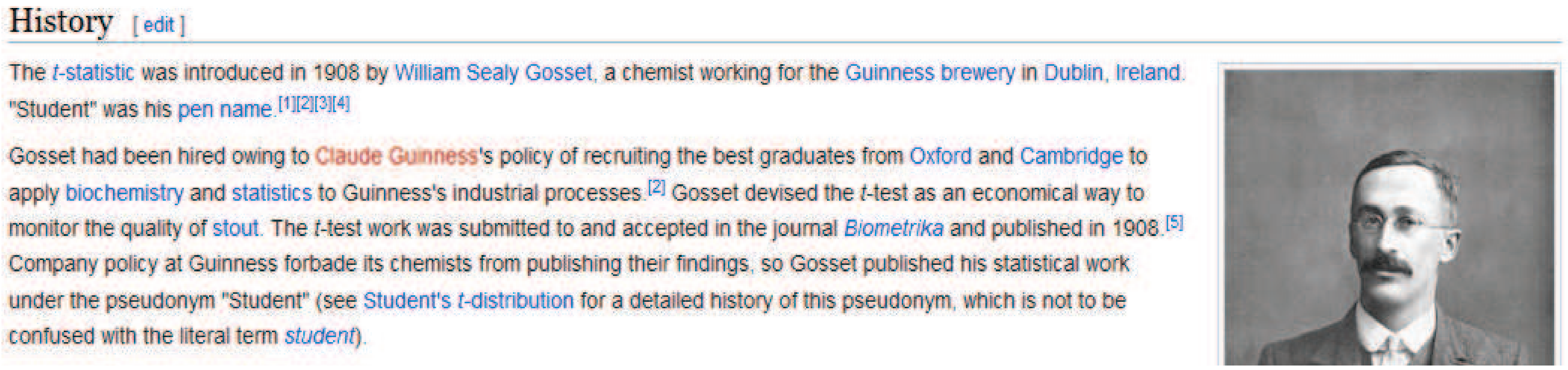
Consequence of CLT for population mean

- We can say things like 95% of sample means fall within 2 $\sigma/\sqrt{n}$ of μ
- We can say things like a sample mean is far from a hypothesized population mean if it is more than 2 SDs away
  - (sample mean – hypothesized mean)/($\sigma/\sqrt{n}$)

t-interval and t-test for one quantitative variable

From the CLT, we know that the distribution of **sample means** is approximately normally distributed…

…but to use the normal distribution for a test of significance or confidence interval we need to specify a mean and SD.



History [edit]

The *t*-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. "Student" was his pen name.[1][2][3][4]

Gosset had been hired owing to Claude Guinness's policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness's industrial processes.[2] Gosset devised the *t*-test as an economical way to monitor the quality of stout. The *t*-test work was submitted to and accepted in the journal *Biometrika* and published in 1908.[5] Company policy at Guinness forbade its chemists from publishing their findings, so Gosset published his statistical work under the pseudonym "Student" (see Student's *t*-distribution for a detailed history of this pseudonym, which is not to be confused with the literal term *student*).

The t-distribution

- With quantitative data, we want to calculate

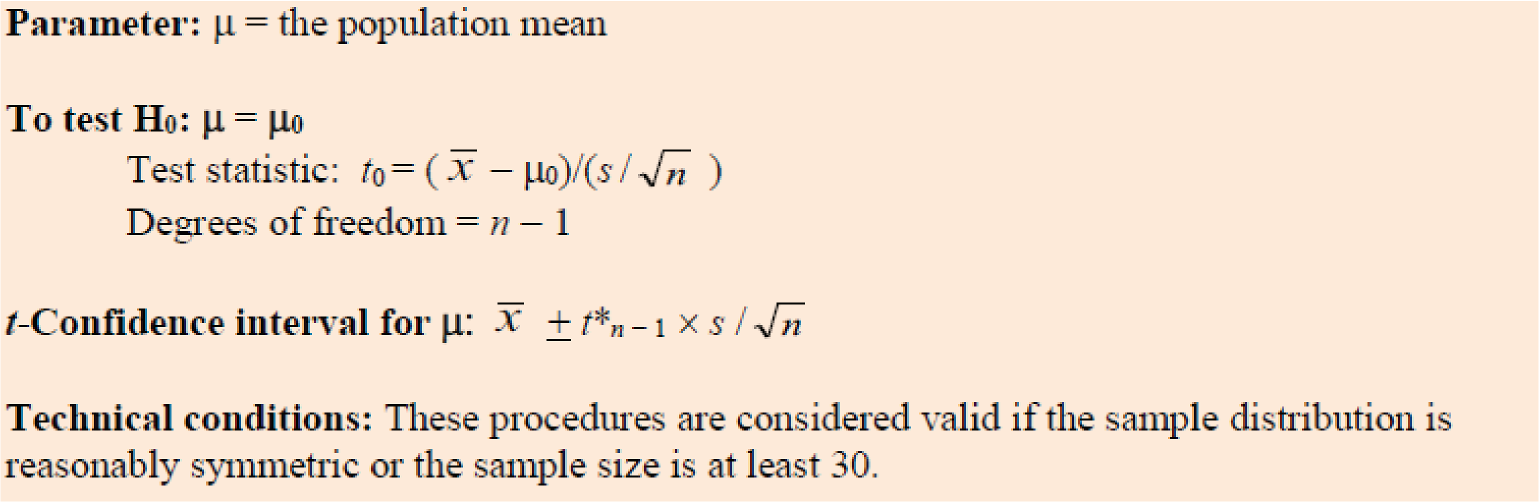
$$SD(\bar{x}) = \sigma / \sqrt{n}$$

- But we don't usually know σ
- But we can calculate the *standard error*

$$SE(\bar{x}) = s / \sqrt{n}$$

- But then our standardized statistic $t = \frac{\bar{x} - \mu_o}{s / \sqrt{n}}$ is better modeled by a *t* distribution (df = *n* -1) than a normal distribution
  - Looks more and more like normal as *n* increases

One sample t-test and t-interval



Parameter: μ = the population mean

To test H_0: $\mu = \mu_0$

Test statistic: $t_0 = (\bar{x} - \mu_0)/(s/\sqrt{n})$

Degrees of freedom = $n - 1$

***t*-Confidence interval for μ:** $\bar{x} \pm t^*_{n-1} \times s/\sqrt{n}$

Technical conditions: These procedures are considered valid if the sample distribution is reasonably symmetric or the sample size is at least 30.