Inference for Numerical Data III

Dajiang Liu @ PHS 525

Feb 23th, 2016

Central Limit Theorem

- The sample mean point estimates $\bar{X}$ is approximately normal.

$$\bar{X} \sim N\left(\mu, se(\bar{X})\right)$$

- The approximation works when:
  - Sample size is "large"
    - A rule of thumb is sample size $n \geq 30$
  - The distribution should not be skewed (i.e. be symmetric)
  - There are no outliers

- The approximation may not be good if any of the above 3 conditions are not met

Sampling Distribution for Different Sample Sizes

Population Distribution does not need to be normal

Sample mean is still normal when sample sizes are large enough

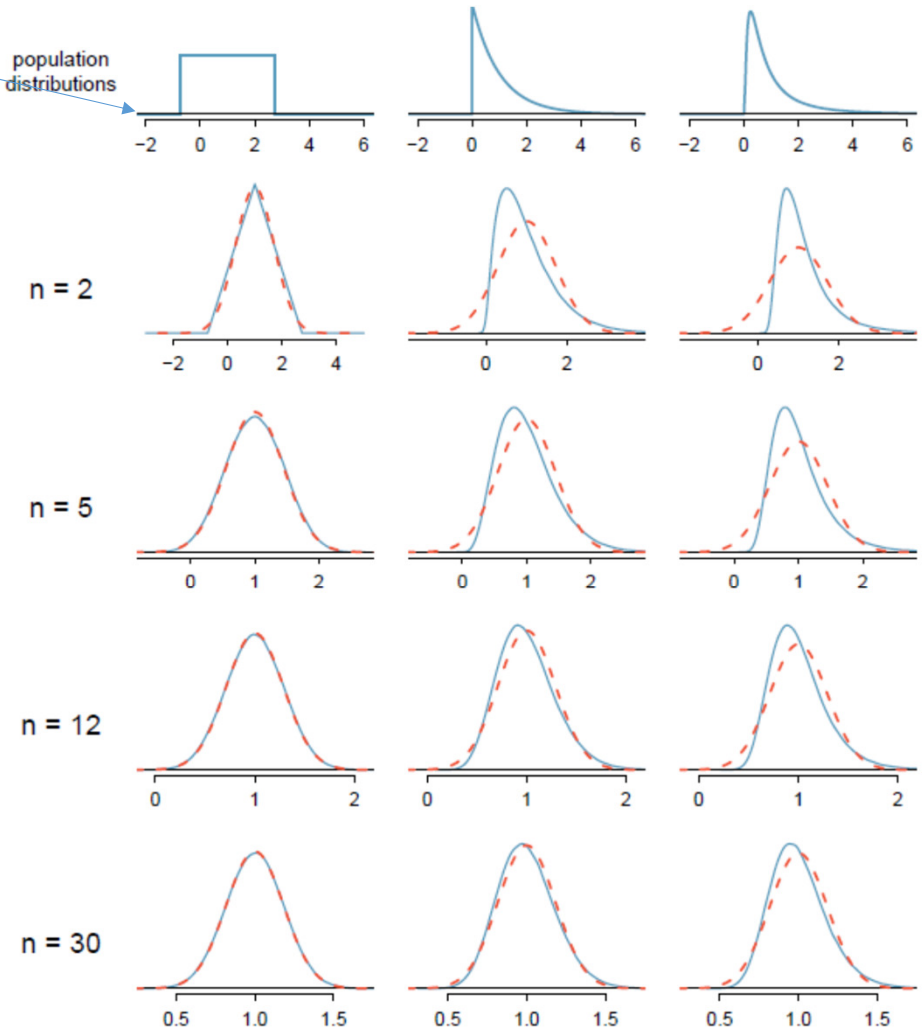
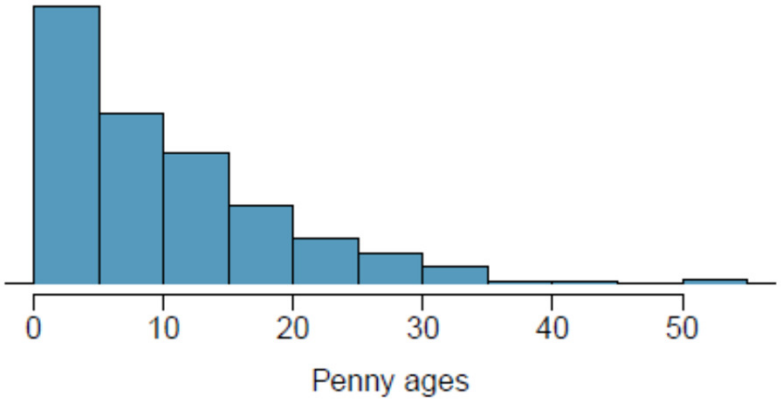


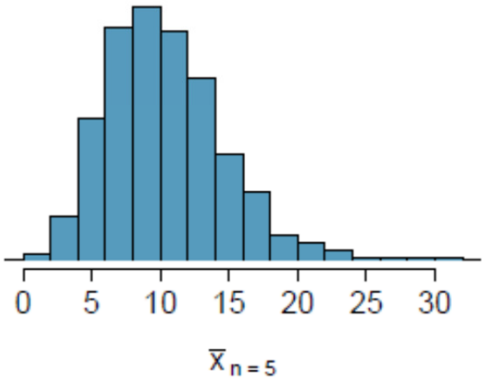
Figure 4.20: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

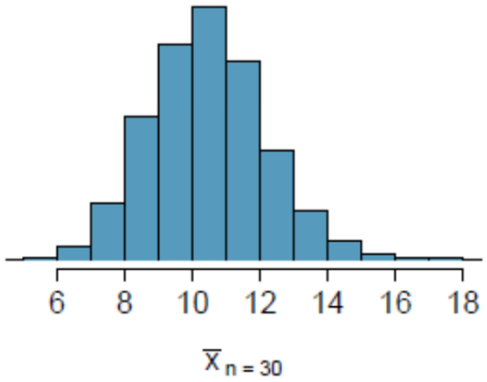
4.33 Ages of pennies, Part I. The histogram below shows the distribution of ages of pennies at a bank.

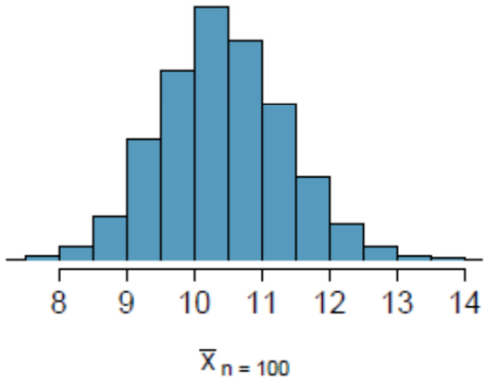
(a) Describe the distribution.

(b) Sampling distributions for means from simple random samples of 5, 30, and 100 pennies is shown in the histograms below. Describe the shapes of these distributions and comment on whether they look like what you would expect to see based on the Central Limit Theorem.







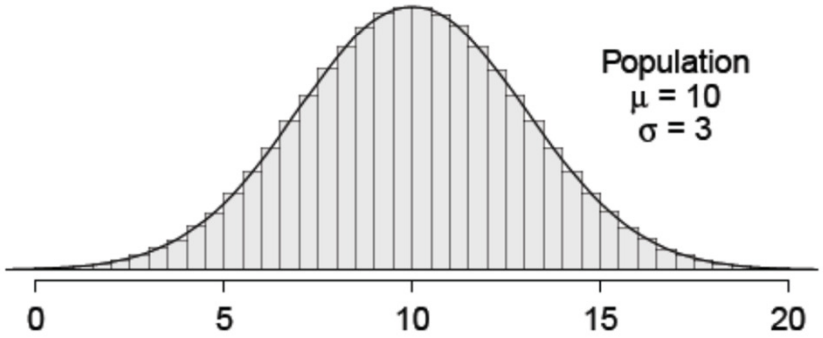


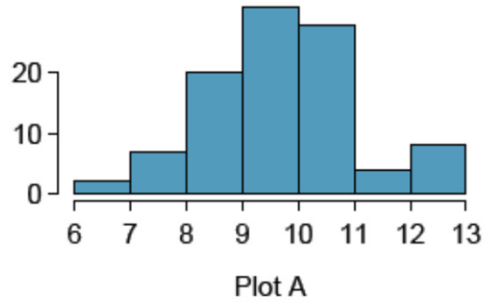
4.33 Answer

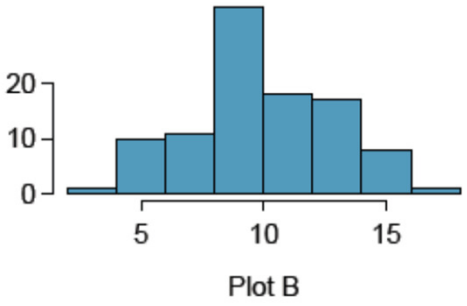
- (a) The distribution is skewed toward smaller values and has several very large outliers

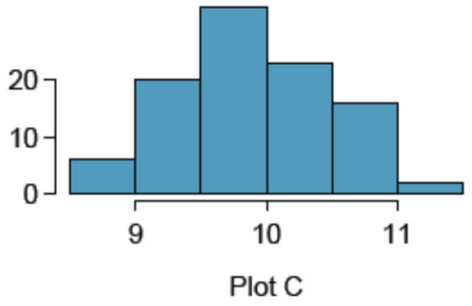
- (b) As sample size gets larger, the distribution of the sample mean estimator behave more like normal distribution. Yet, there are still heavy upper tails, possibly due to the influence of the outliers with large values.

4.35 Identify distributions, Part I. Four plots are presented below. The plot at the top is a distribution for a population. The mean is 10 and the standard deviation is 3. Also shown below is a distribution of (1) a single random sample of 100 values from this population, (2) a distribution of 100 sample means from random samples with size 5, and (3) a distribution of 100 sample means from random samples with size 25. Determine which plot (A, B, or C) is which and explain your reasoning.









4.35 Answer

- (1) -> (b)
- (2) -> (a)
- (3) -> (c)
- The key is to examine the standard error. The sample mean from larger samples has the smallest standard errors.

One Sample Means with t-distribution

- Central Limit Theorem requires large sample sizes
  - In large samples, sample mean estimate is more likely to be normally distributed
  - In large samples, the sample mean estimate tend to have smaller standard deviation

Yet:

- In many cases, large samples can be hard to attain

- t-distribution can be a helpful alternative for small sample inference

The Normality Condition – Modified

- Central limit theorem modified:
  - The sampling distribution for the mean is nearly normal when the sample observations are independent and come **from a nearly normal distribution**.
  - Important to note:
    - The CLT modified does not put constraint on the sample size
    - The CLT modified does require that population distribution is nearly normal
      - Original CLT does not require population distribution be normal
  - Even for sample sizes, CLT modified holds.

Degrees of Freedom (df)

- Degrees of freedom measure the shape of the distribution

- The larger the df, the more closely the t-distribution resembles the normal distribution

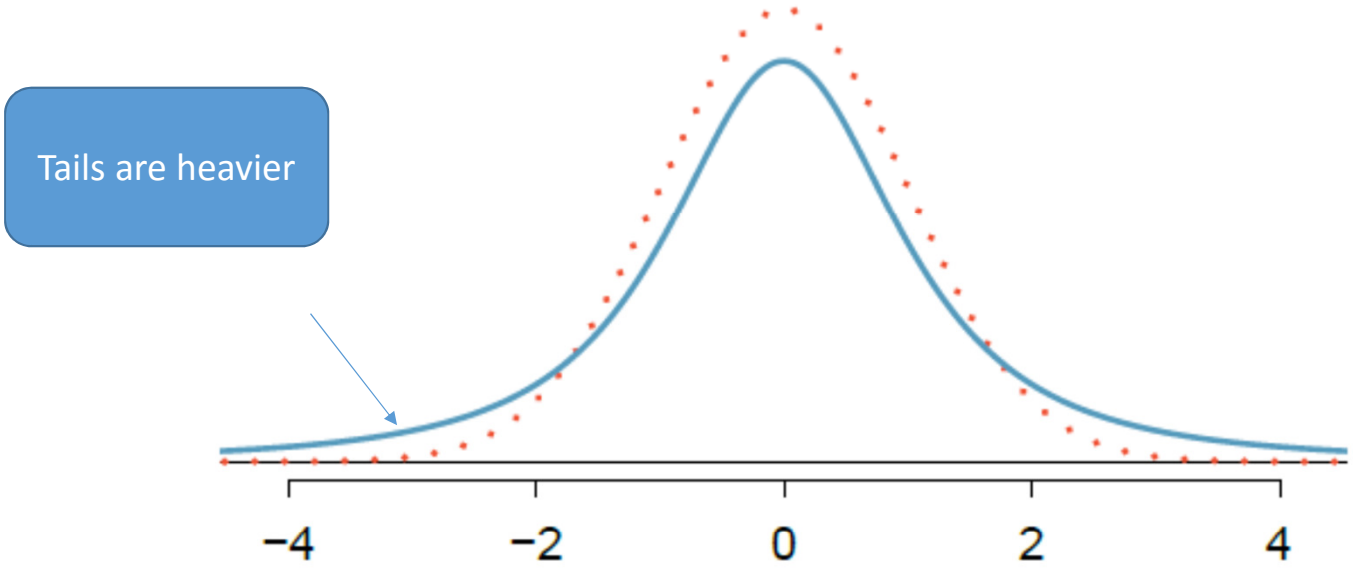


Figure 5.10: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

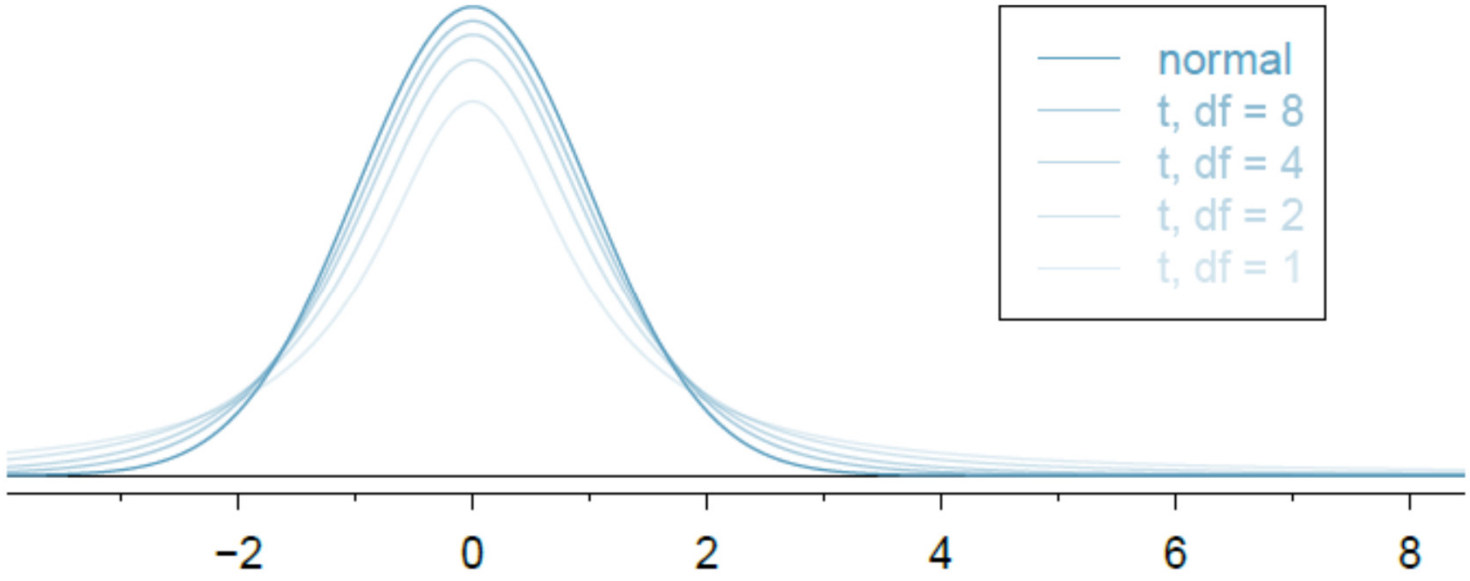


Figure 5.11: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

Use t-distribution to Obtain Confidence Interval

- Confidence intervals obtained using t-distribution can be more accurate

- Procedures for obtaining t-distribution based confidence interval
  - Obtain sample mean point estimate $\bar{X}$
  - Obtain sample standard deviation σ
  - Obtain standard error for the sample mean point estimate
    - $se(\bar{X}) = \sigma/\sqrt{n}$
  - Confidence interval is obtained by
  - $\bar{X} - t_{df=n-1} \times se(\bar{X}) \leq \mu \leq \bar{X} + t_{df=n-1} \times se(\bar{X})$
  - $t_{df=n-1}$ is the critical t-value

5.15 Identify the critical t. An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t value ($t^\star$) for the given sample size and confidence level.

(a) $n = 6$, CL $= 90\%$

(b) $n = 21$, CL $= 98\%$

(c) $n = 29$, CL $= 95\%$

(d) $n = 12$, CL $= 99\%$

Example: What is the normal and t-distribution based confidence interval??

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 5.16: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Example: What is the normal and t-distribution based 95%-confidence interval??

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 5.16: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Answer:

$se(\bar{X}) = \frac{2.3}{\sqrt{19}}$ =0.53

Normal confidence interval equals to (3.36,5.44)

t confidence interval equals to (3.29,5.51)

Exercise 5.20 The FDA's webpage provides some data on mercury content of fish.[14] Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[15]

Example 5.21 Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Exercise 5.20. If we are to use the t distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^\star_{df}$.

Hypothesis Testing with t-Distribution

- T statistic
- For a sample of size n,
  - Estimate sample mean $\bar{X}$ and standard deviation $se(\bar{X})$
  - To test the hypothesis $H_0: \mu = \mu_0$ v.s. $H_A: \mu > \mu_0$
  - A t-statistic can be calculated

$$T = \frac{(\bar{X} - \mu_0)}{se(\bar{X})}$$

The p-value can be assessed by $\Pr(T^* > T)$, where T^* is a random variable with distribution $t_{df=n-1}$

5.18 Find the p-value. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given set of hypotheses and T test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

(a) $H_A : \mu > \mu_0$, $n = 11$, $T = 1.91$

(b) $H_A : \mu < \mu_0$, $n = 17$, $T = -3.45$

(c) $H_A : \mu \neq \mu_0$, $n = 7$, $T = 0.83$

(d) $H_A : \mu > \mu_0$, $n = 28$, $T = 2.13$

5.18 **Find the p-value.** An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given set of hypotheses and T test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

(a) $H_A : \mu > \mu_0$, $n = 11$, $T = 1.91$

(b) $H_A : \mu < \mu_0$, $n = 17$, $T = -3.45$

(c) $H_A : \mu \neq \mu_0$, $n = 7$, $T = 0.83$

(d) $H_A : \mu > \mu_0$, $n = 28$, $T = 2.13$

Answer and R command:

(a). pt(1.91,df=10,lower.tail=FALSE)

[1] 0.04260244

(b). 2*pt(0.83,df=6,lower.tail=FALSE)

[1] 0.4383084

(c). pt(-3.45,df=16,lower.tail=TRUE)

[1] 0.001646786

(d). pt(2.13,df=28,lower.tail=FALSE)

[1] 0.02104844

5.19 Sleep habits of New Yorkers. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. Do these data provide strong evidence that New Yorkers sleep less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(a) Write the hypotheses in symbols and in words.

(b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

(c) Find and interpret the p-value in this context. Drawing a picture may be helpful.

(d) What is the conclusion of the hypothesis test?

(e) If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

Answer 5.19

- (a). $H_0: \mu = 8$ v.s. $H_A: \mu < 8$
- (b). $T = \frac{7.73-8}{0.77/\sqrt{25}} = -\frac{0.27}{0.77} \times 5 = -1.75$
- (c). P-value 0.046
- (d). Reject the null hypothesis at $\alpha = 0.05$
- (e). (7.47,7.99)