Lecture 4

Confidence Intervals

Lecture Summary

- Last lecture, we talked about summary statistics and how “good” they were in estimating the parameters
  - Risk, bias, and variance
  - Sampling distribution

- Another quantitative measure of how “good” the statistic is called **confidence intervals (CI)**

- CIs provide an interval of certainty about the parameter

Introduction

- Up to now, we obtained **point estimates** for parameters from the sample $X_1, \dots, X_n$
  - Examples: sample mean, sample variance, sample median, sample quantile, IQR, etc.
  - They are called **point estimates** because they provide one single point/value/estimate about the parameter
  - Mathematically: $T(X_1, \dots, X_n) \rightarrow$ a single point!

- However, suppose we want a range of possible estimates for the parameter, an interval estimate like $[L, U]$
  - Mathematically: $L(X_1, \dots, X_n)$ and $U(X_1, \dots, X_n)$

Two-Sided Confidence Intervals

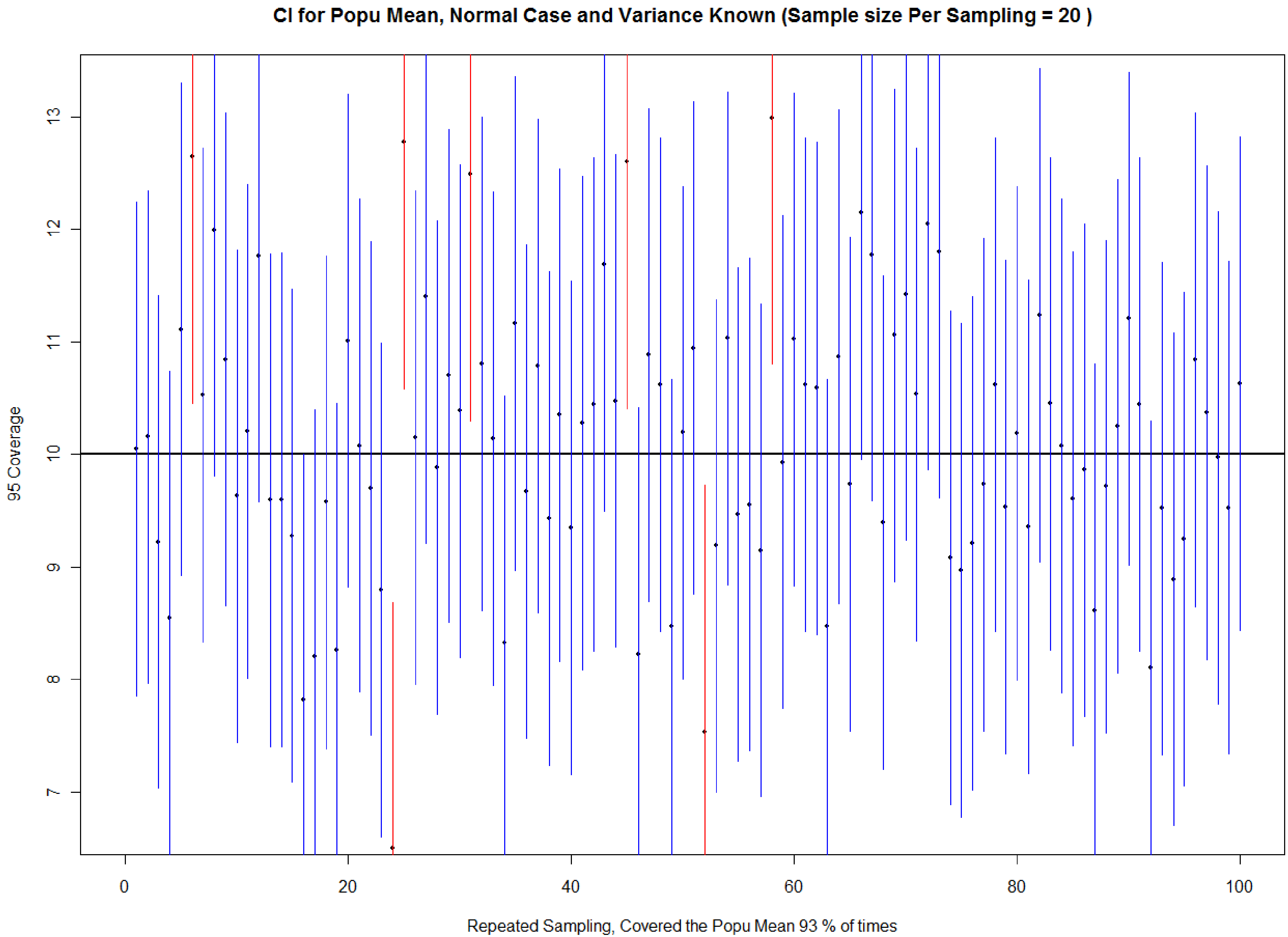
- <u>Data/Sample</u>: $(X_1, \dots, X_n) \sim F_\theta$
  - θ is the parameter

Confidence Level

- **Two-Sided Confidence Intervals**: A $\boxed{\alpha}$-confidence interval is a random interval, $[L, U]$, from the sample where the following holds

$$P(L \leq \theta \leq U) \geq 1 - \alpha$$

  - <u>Interpretation</u>: the probability of the interval covering the parameter must exceed $1 - \alpha$
  - It is NOT the probability of the parameter being inside the interval!!! Why?


CI for Popu Mean, Normal Case and Variance Known (Sample size Per Sampling = 20)
95 Coverage
Repeated Sampling, Covered the Popu Mean 93 % of times
7
8
9
10
11
12
13
0
20
40
60
80
100

Comments about CIs

- Pop quiz 1: What is the confidence level, α, for $[-\infty, \infty]$ confidence interval?
  - Thus, for any level α, $[-\infty, \infty]$ CI would be a valid (but terrible) CI

- Pop quiz 2: What is the confidence level, α, for $[a, a]$ CI where a is any number?

- Pop quiz 3: Suppose you have two confidence intervals $[L_1, U_1]$ and $[L_2, U_2]$. If the first CI is shorter than the second, what does this imply?
  - If α is the same for both intervals, what would this imply about the short interval (in comparison to the longer interval)?

- Main point: given some confidence level α, you want to obtain the shortest CI

CIs for Population Mean

- Case 1: If the population is Normal and σ^2 is known

  CI: $\bar{X} \pm Z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$

  Hint: Use sampling distribution of $\bar{X}$

- Case 2: If the population is not Normal and σ^2 is known

  Approximate CI: $\bar{X} \pm Z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$

  Hint: Use CLT of $\bar{X}$

What if the variance, σ^2, is unknown?

t Distribution

- Formal Definition: A random variable X has a t-distribution with n degrees of freedom, denoted as t_n, with the probability density function

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

where $\Gamma()$ is a gamma function

- Useful Definition: Consider the following random variable

You will prove the relation between the two in the homework

$$X = \frac{Z}{\sqrt{\frac{V}{n}}}$$

Notice that you can transform $\bar{X} - \mu$ into a standard Normal

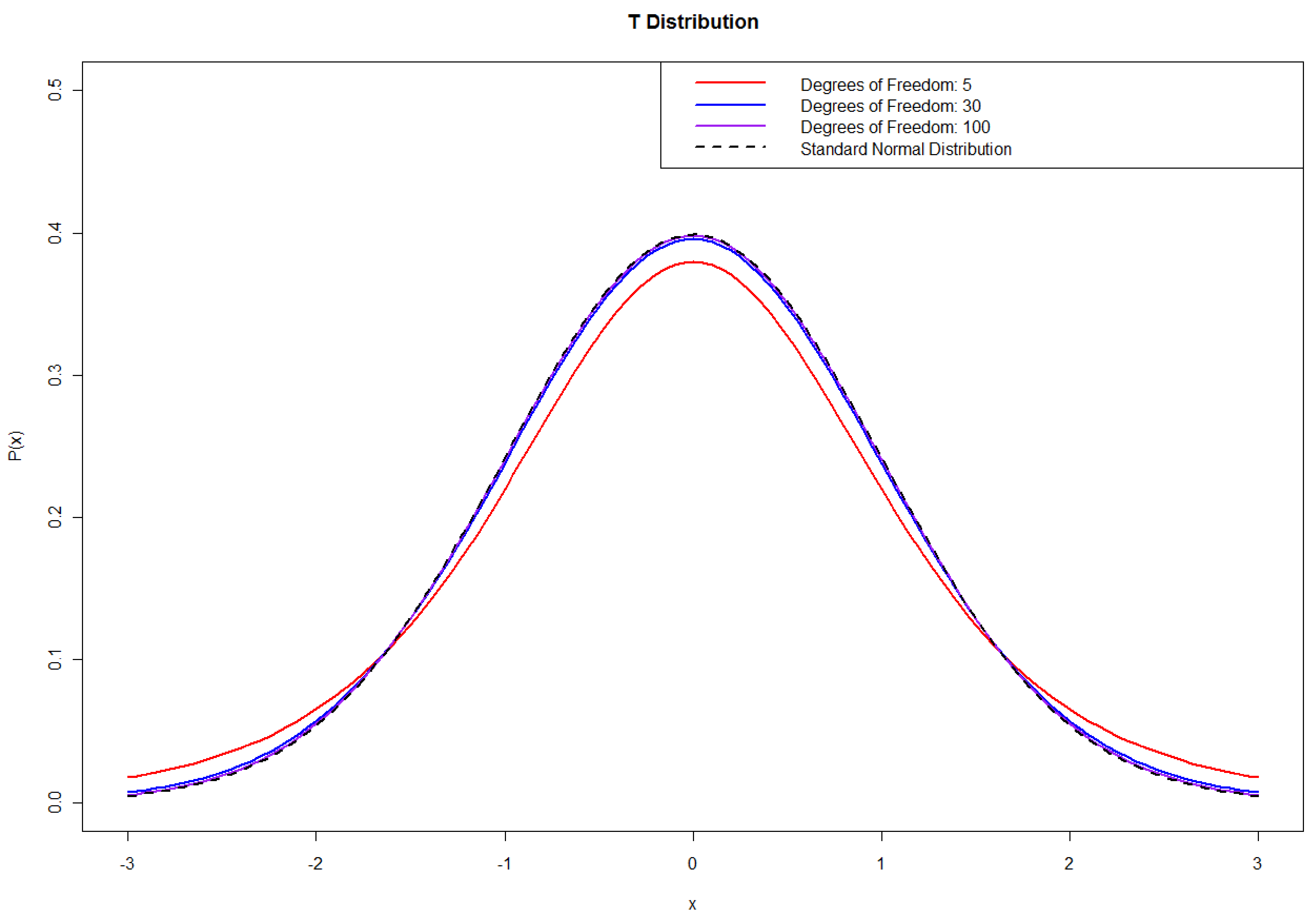
where $Z \sim N(0,1)$, $V \sim \chi_n^2$ and Z and V are independent.

Then, $X \sim t_n$

- "Quick and Dirty" Definition: If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, i.i.d., then

$$\frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

From lecture 3, $\hat{\sigma}^2$ is χ_{n-1}^2, with some constant multipliers



T Distribution
Degrees of Freedom: 5
Degrees of Freedom: 30
Degrees of Freedom: 100
Standard Normal Distribution
P(x)
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3
x

Property of the t Distribution

- The t-distribution has a fatter tail than the normal distribution (see picture from previous slide)
  - Consequences: The "tail" probabilities for the t distribution is bigger than that from the normal distribution!

- If the degrees of freedom goes to ∞, then

$$\lim_{n \to \infty} t_n \to N(0,1)$$

  Proof: CLT!

- This means that with large sample size (n), we can approximate t_n with a standard normal distribution

$$P(t_n \leq x) \approx P(Z \leq x)$$

  for large n
  - General rule of thumb for how large n should be: $n \geq 30$

CIs for Population Mean

- Case 3: If the population is Normal and variance is unknown

  CI: $\bar{X} \pm t_{\left(1-\frac{\alpha}{2}\right)} \frac{\hat{\sigma}}{\sqrt{n}}$

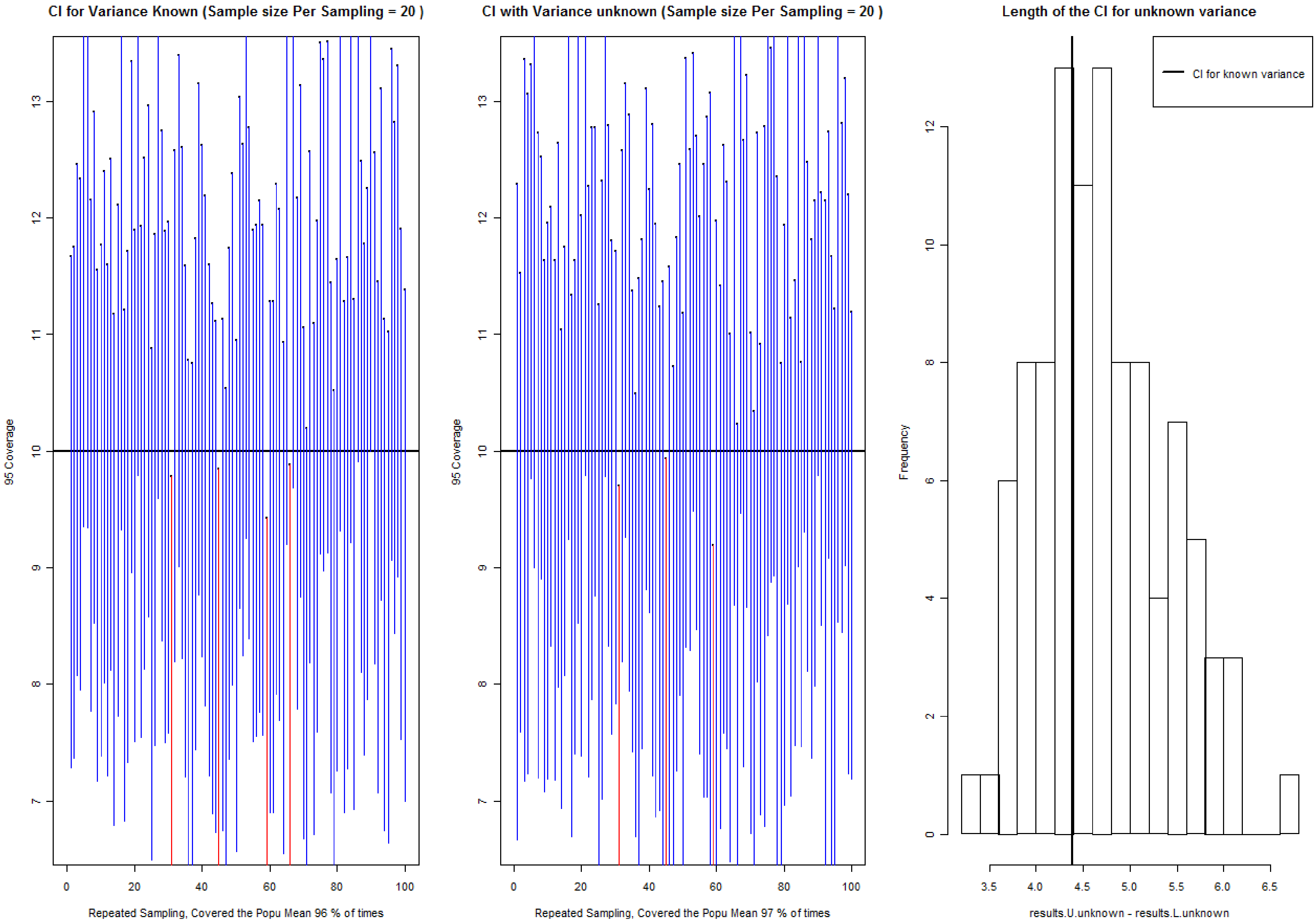
  Hint: Use the "quick and dirty" version of the t-distribution

- Case 4: If the population is not Normal and variance is unknown (i.e. the "realistic" scenario)

  Approximate CI: $\bar{X} \pm z_{\left(1-\frac{\alpha}{2}\right)} \frac{\hat{\sigma}}{\sqrt{n}}$

  Hint: Use CLT!

  – Demo in class



CI for Variance Known (Sample size Per Sampling = 20)
95 Coverage
Repeated Sampling, Covered the Popu Mean 96 % of times
CI with Variance unknown (Sample size Per Sampling = 20)
95 Coverage
Repeated Sampling, Covered the Popu Mean 97 % of times
Length of the CI for unknown variance
CI for known variance
Frequency
results.U.unknown - results.L.unknown

Summary of CIs for the Population Mean

Fixed width CI

Scenarios	CI	Derivation
1) Population is Normal 2) Variance is known	$\bar{X} \pm Z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$	Use sampling distribution for $\bar{X}$
1) Population is not Normal 2) Variance is known	$\bar{X} \pm Z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}$	Approximate CI, use CLT
1) Population is Normal 2) Variance is unknown	$\bar{X} \pm t_{\left(1-\frac{\alpha}{2}\right)} \frac{\hat{\sigma}}{\sqrt{n}}$	Use the t distribution
1) Population is not Normal 2) Variance is unknown	$\bar{X} \pm z_{\left(1-\frac{\alpha}{2}\right)} \frac{\hat{\sigma}}{\sqrt{n}}$	Approximate CI, use CLT

Variable width CI

CIs for Population Variance

- Case I: If the population is Normal and all parameters are unknown

  $$[(n-1)\hat{\sigma}^2/\chi^2_{n-1\left(1-\frac{\alpha}{2}\right)}, (n-1)\hat{\sigma}^2/\chi^2_{n-1\left(\frac{\alpha}{2}\right)}]$$

  – Hint: Use the sampling distribution for $\hat{\sigma}^2$

- Case II: (Homework question) If the population is Normal and the population mean is known.

  – Hint: Use $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ and the sampling distribution related to it!

Lecture Summary

- Another quantitative measure of how "good" the statistic is called **confidence intervals (CI)**
- CIs provide an interval of certainty about the parameter
- We derived results for the population mean and the population variance, under various assumptions about the population
  - Normal vs. not Normal
  - known variance vs. unknown variance

Extra Slides

One-Sided Confidence Intervals

- **One-Sided Confidence Intervals**: A α-confidence interval is a random interval, $[L, \infty]$, from the sample where the following holds
$$P(L \leq \theta) \geq 1 - \alpha$$

- **One-Sided Confidence Intervals**: A α-confidence interval is a random interval, $[-\infty, U]$, from the sample where the following holds
$$P(\theta \leq U) \geq 1 - \alpha$$