ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

INFERENCE FOR NUMERICAL DATA

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

- In previous modules, we considered inference in the following settings, all involving ***categorical data***:
    - A single proportion
    - Difference of two proportions
    - Multiple groups
- We
    - Constructed confidence intervals
    - Conducted hypothesis tests
- Here, we consider inference, in the setting of **numerical data.**

Introduction

- Here, we consider inference, in the setting of **numerical data.** We will focus on:
  - A single mean
  - Paired data
  - Difference of two means
  - Many means
- We will construct
  - Confidence intervals
  - Conduct hypothesis tests

One-Sample Means and the t-Distribution

Categorical Data

- Sample proportion: $\hat{p}$
  Population proportion: p
- Modeled $\hat{p}$ using normal distribution – centered at p and with $SE = \sqrt{\dfrac{p(1-p)}{n}}$.
- Used properties of the normal distribution to construct confidence intervals and conduct hypothesis tests.

Numerical Data

- Sample mean: $\bar{x}$
  Population mean: μ
- **Will model $\bar{x}$ using t-distribution (and a single parameter, the degrees of freedom df)**
- Will use properties of the t-distribution distribution to construct confidence intervals and conduct hypothesis tests

Why the t-Distribution?

- **Central limit theorem (for sample mean):**
  If our sample consists of n independent observations from a population with mean μ and standard deviation σ, and n is large enough,

  then the sampling distribution of $\bar{x}$ is nearly normal with

  mean = μ $\qquad$ $SE = \dfrac{\sigma}{\sqrt{n}}.$

- Two issues to consider:
- Conditions under which the CLT approximation can be safely used.
- In practice, we don't know σ, so we must estimate it. As our estimation is imperfect we use a new distribution: the t-distribution to resolve this issue.

Conditions Needed to Apply the CLT

- As in the categorical data setting, we need two conditions to be satisfied to apply the CLT for a sample mean
  - Independence: The sample observations must be independent. For example, this happens if our sample is a random sample from a large population.
  - Normality: If the n is small, we require that the sample observations come from a normally distributed population. This condition can be relaxed as n increases.
- **Rules of thumb for normality:**
  - $n < 30$: If there are no outliers in the data, we assume normality of the data, which implies normality of $\bar{x}$.
  - $n \geq 30$: If there are no extreme outliers in the data, we assume normality of $\bar{x}$ even if distribution of the observations is not.

Estimating σ, Introducing the t-Distribution

- In practice, we don't know the population mean μ or standard deviation σ .
- As in the categorical data case, we will use the sample value as a proxy for the population value.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

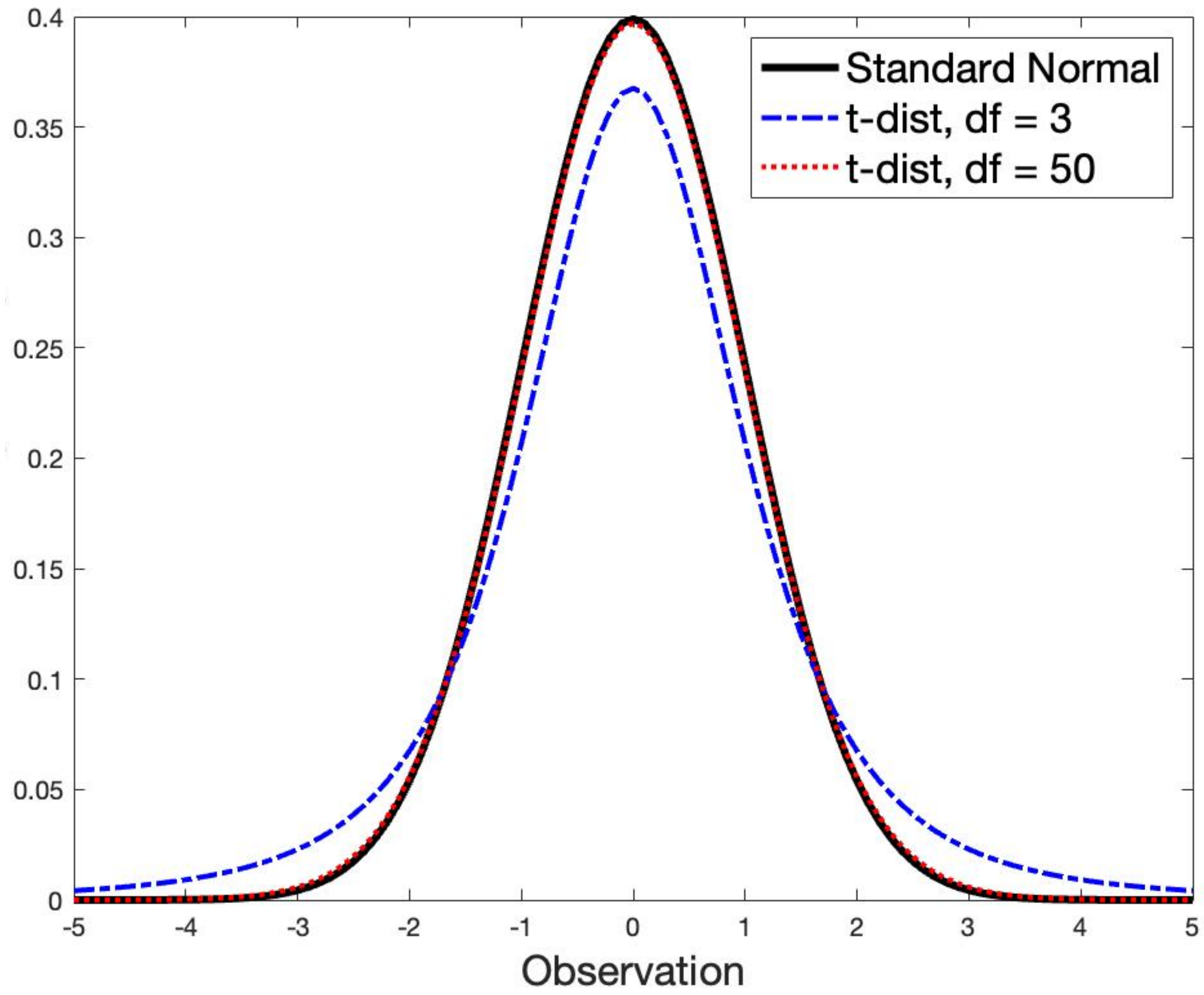
- **t-distribution:**
  - Always centered at 0.
  - Parametrized by a single parameter: the degrees of freedom df.
  - In general $df = n - 1.$

Estimating σ, Introducing the t-Distribution

- In practice, we don't know the population mean μ or standard deviation σ .
- As in the categorical data case, we will use the sample value as a proxy for the population value.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

- **t-distribution:**
  - Always centered at 0.
  - Parametrized by a single parameter: the degrees of freedom **df.**
  - In general $df = n - 1.$



The larger the degrees of freedom, the more closely the t-distribution approximates the standard normal.