ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DISTRIBUTION OF RANDOM VARIABLES:

OTHER DISTRIBUTIONS

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

The Chi-Squared Distribution, T-Distribution, and F-Distribution

- We will soon see that in statistical inference, there are procedures that — loosely speaking — involve (among other things):

- Using the data (i.e., random samples drawn from a distribution) to calculate a test statistic (a function of the samples) which can be thought of as being random variable itself, *albeit from a potentially different distribution.*

- Under some assumptions on the distributions of the data, understanding the resulting distribution of the test-statistic.

- Calculating the probability that a random variable drawn from that distribution of the test-statistic is as extreme as the observed test statistic, e.g., $\mathbb{P}(X \geq \text{test statistic})$.

An Illustrative Example

- **Example:** Suppose you want to understand whether a coin is fair or not. So you flip the coin $n = 100$ times, and your record the outcome of each trial. In particular, you know the total number, k, of heads (successes) that result. **Say you got $k = 60$ heads.**
- You might decide that your test-statistic is the Z-score associated with k, *assuming that the coin is fair, so that* $\mu = 50$, $\sigma = \sqrt{100 \times 0.5(1 - 0.5)} = 5$.
- In other words you calculate $Z_n = \dfrac{k - 50}{5}$. In our case $Z_n = 2$.
- By the central limit theorem, we can approximate the distribution of Z by $\mathcal{N}(0,1)$.
- **Now, we can ask:** If the coin was fair, hence $Z \sim \mathcal{N}(0,1)$, what would the probability of observing data as extreme as the test-statistic, i.e., $\mathbb{P}(Z \geq Z_n)$.
- In this case, we can calculate this probability as ≈ 0.0228

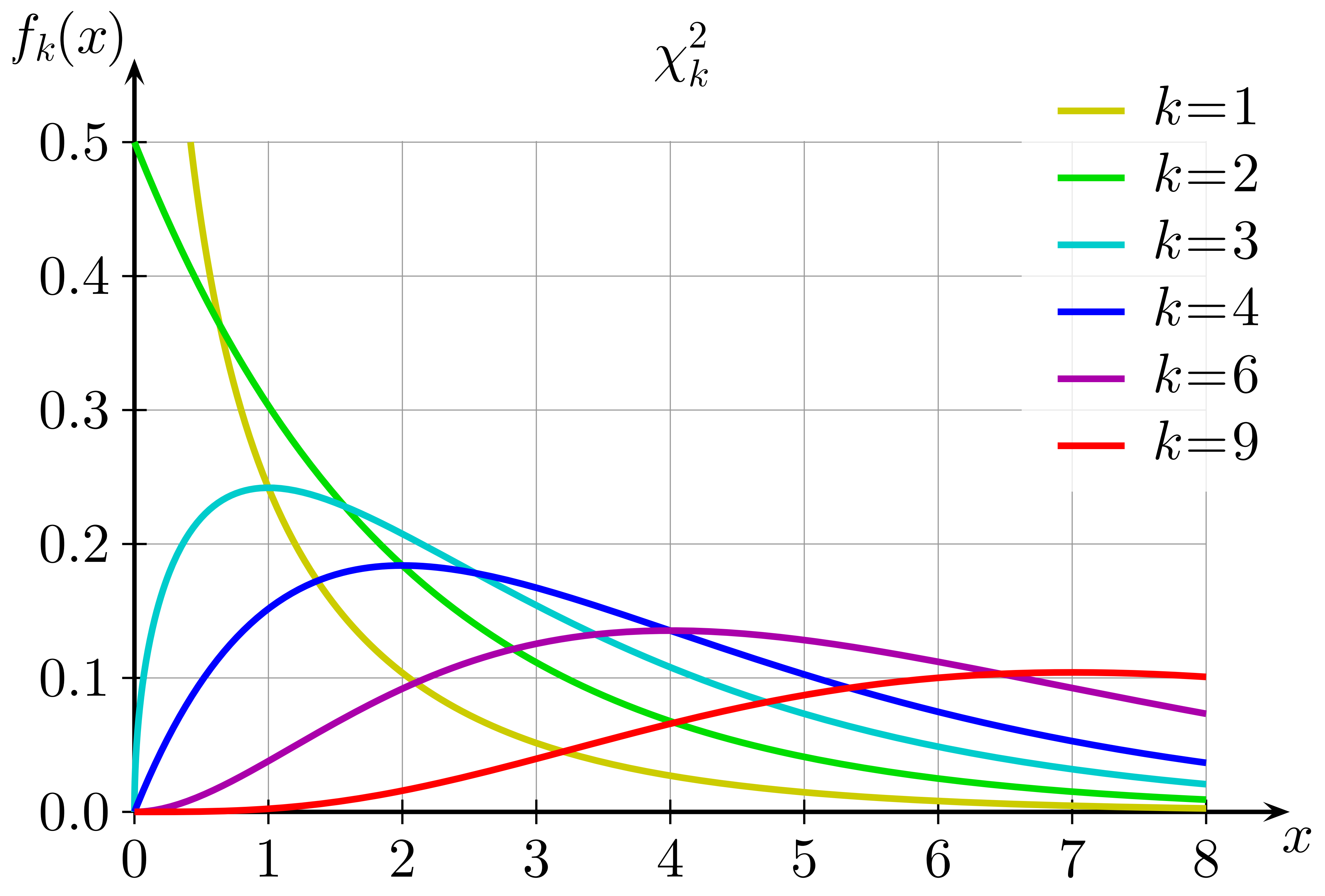
The Chi-Squared Distribution

- Under different statistical scenarios, different test-statistics are appropriate, and they are associated with different distributions.

- **Example:** When trying to assess how well given data fits a particular distribution, the chi-squared test statistic is useful:

$$Q = \sum_{i=1}^{k} Z_i^2$$

- When $Z_i \sim \mathcal{N}(0,1)$ then Q is distributed according to the χ^2-distribution, with k degrees of freedom.



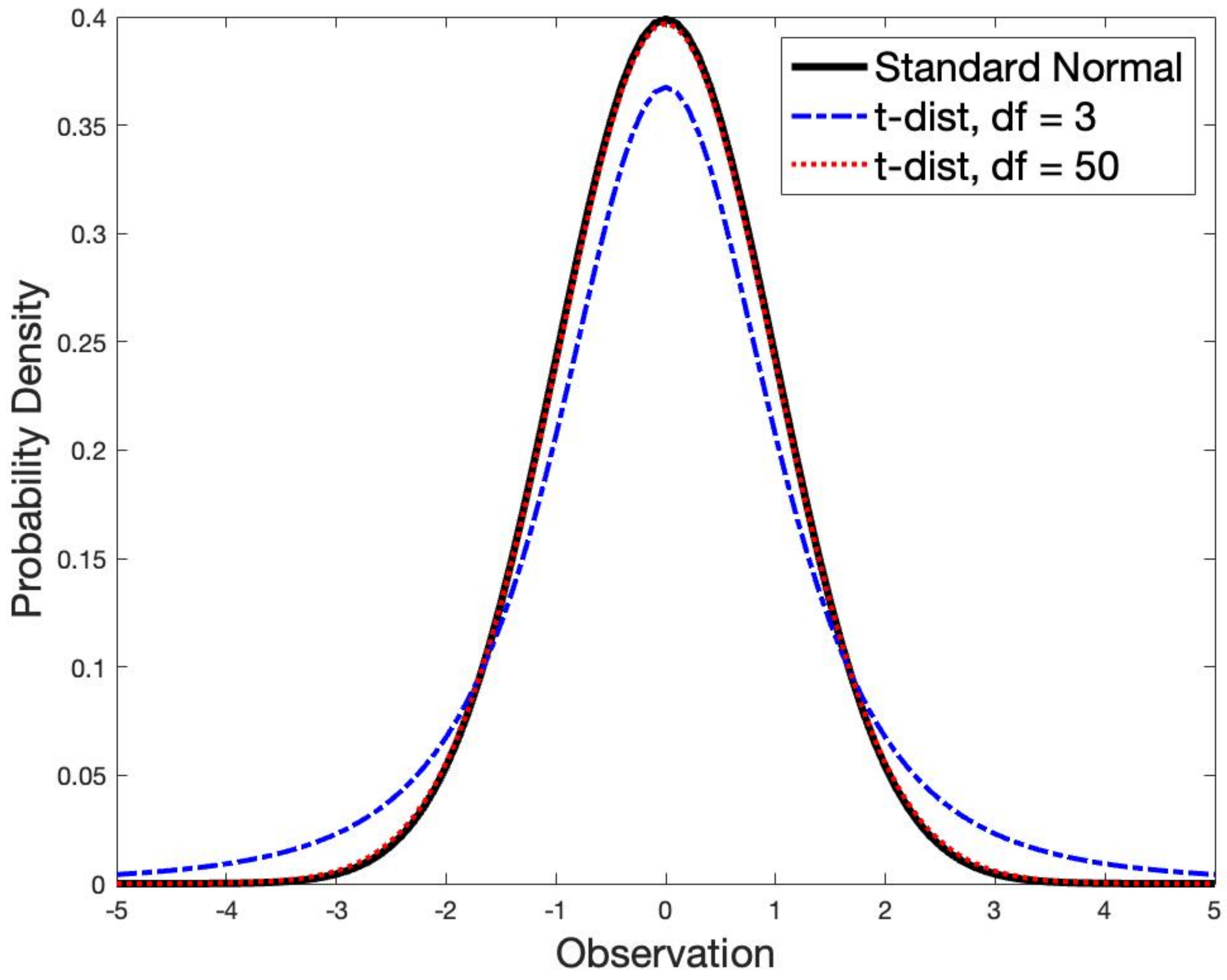
The probability density function of chi-squared distributions with k degrees of freedom.

The T-Distribution

- There are scenarios where the true variance is not known, but estimated from the data – i.e., we use the sample variance
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
in place of σ^2.

- While the random variable $\dfrac{\bar{X} - X_i}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, the random variable $\dfrac{\bar{X} - X_i}{s/\sqrt{n}}$ is ***not normally distributed.***

- Instead $\dfrac{\bar{X} - X_i}{s/\sqrt{n}}$ is distributed according to the t-distribution, with $n-1$ degrees of freedom.



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

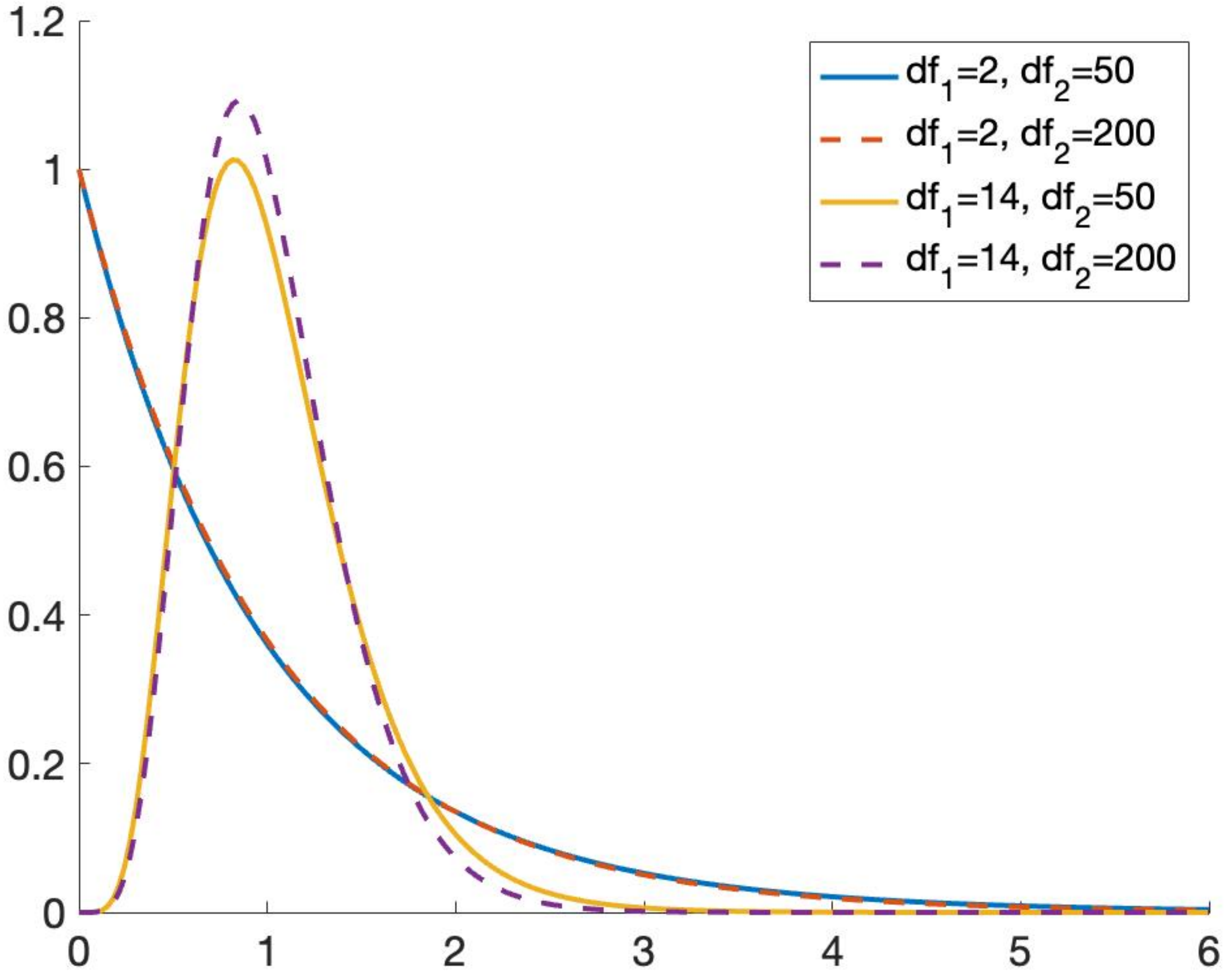
The F-Distribution

- Later (when we consider ANOVAs), we'll see scenarios where the test statistic is a ratio

$$W = \frac{\frac{S_1}{d_1}}{\frac{S_2}{d_2}}$$

  where $S_1 \sim \chi^2$-distribution with d_1 degrees of freedom and $S_2 \sim \chi^2$-distribution with d_2 degrees of freedom.

- In this case, W is distributed according to the **F-distribution,** parametrized by two parameters d_1 and d_2.



The Chi-Squared Distribution, T-Distribution, and F-Distribution

- We will see all these distributions when we consider hypothesis testing.

- We'll need to calculate tale probabilities under these distributions.

- These tale probabilities don't have closed form expressions — *just like tale probabilities don't have closed form expressions even for* $\mathcal{N}(0,1)$.

- We'll have to calculate the relevant tail probabilities using software or tables.