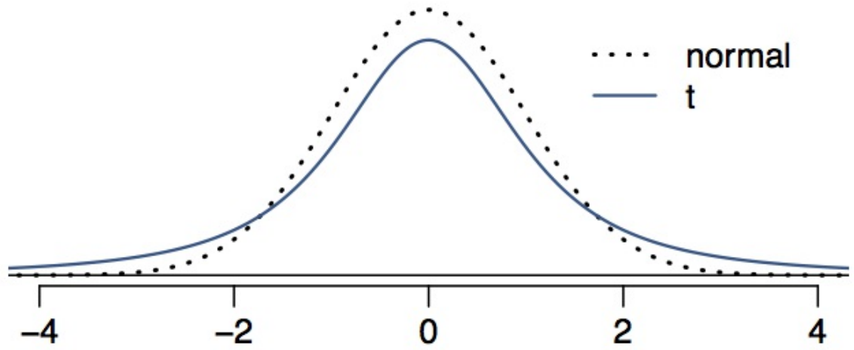
INFERENCE FOR NUMERICAL DATA

t-distribution

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

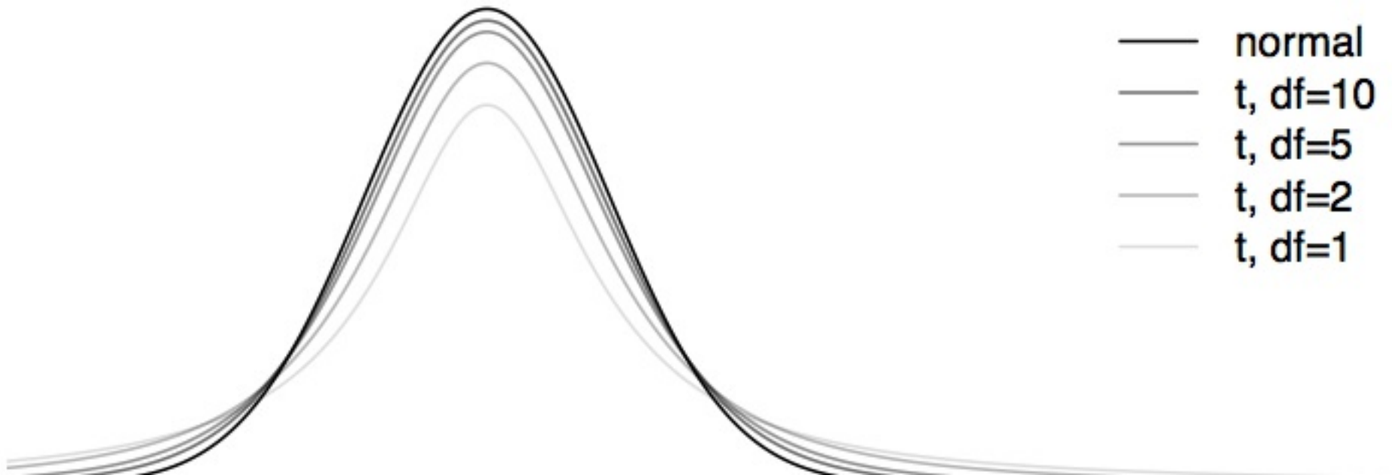
t-distribution

- σ is unknown (almost always) $\rightarrow$ $\bar{x} \sim$ t-distribution
- t-distribution is bell shaped but has thicker tails the normal
- Observations more likely to fall beyond 2 SDs from the mean



Shape of the t-distribution

- Always centered at 0
- Has one parameter: **degrees of freedom** (df) - determines thickness of tails
    - As df increases, the t-distribution approaches the normal distribution



INFERENCE FOR NUMERICAL DATA

Let's practice!

INFERENCE FOR NUMERICAL DATA

Estimating with the t-interval

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

Quantifying variability of sample means

Suppose among a random sample of 100 people 13 are left handed. If you were to select another random sample of 100, would you be surprised if only 12 are left handed? What about 15? Or 30? Or 1 or 90?

Ways to quantify the variability of the sample mean:

- Simulate with bootstrapping
- Approximate with Central Limit Theorem

Central Limit Theorem

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

- SE (standard error) = standard deviation of the sampling distribution
- σ unknown:
  - $SE = \frac{s}{\sqrt{n}}$
  - Use $t_{df=n-1}$ for inference for a mean
- Only true if certain conditions are satisfied...

Conditions

1. Independent observations: Hard to check, but...
    - random sampling / assignment
    - if sampling without replacement, n < 10% of population
2. Sample size / skew: The more skewed the original population, the larger the sample size should be.

Confidence interval for a mean

Estimate the average number of days Americans work extra hours beyond their usual schedule (variable: moredays) using data from the 2010 General Social Survey (data: gss).

Confidence interval for a mean

Estimate the average number of days Americans work extra hours beyond their usual schedule (variable: moredays) using data from the 2010 General Social Survey (data: gss).

```
t.test(gss$moredays, conf.level = 0.95)

    One Sample t-test

data:  gss$moredays
t = 25.628, df = 1146, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.273367 6.147732
sample estimates:
mean of x
 5.710549
```

INFERENCE FOR NUMERICAL DATA

Let's practice!

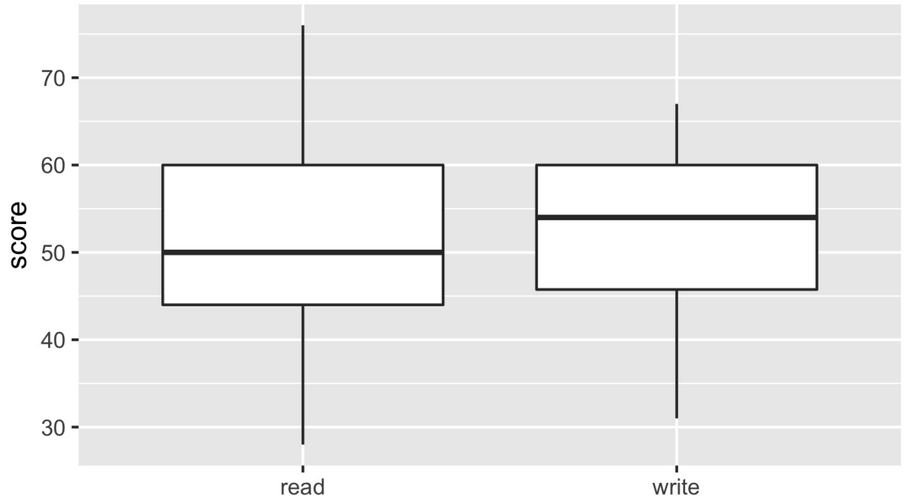
INFERENCE FOR NUMERICAL DATA

t-interval for paired data

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

High School and Beyond

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and writing test. At a first glance, how are the distributions of reading and writing scores similar? How are they different?



Independent scores?

Can reading and writing scores for a given student student assumed to be independent of each other?

Probably not!

Analyzing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be paired.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations: diff = read − write.

student	read	write	diff
1	57	52	5
2	68	59	9
3	44	33	11
...	...	...	...
200	63	65	-2

Estimating the mean difference in paired data

Construct a 95% confidence interval for the mean difference between the average reading and writing scores.

Estimating the mean difference in paired data

Construct a 95% confidence interval for the mean difference between the average reading and writing scores.

```
t.test(hsb2$diff, conf.level = 0.95)
```

Estimating the mean difference in paired data

Construct a 95% confidence interval for the mean difference between the average reading and writing scores.

```
t.test(hsb2$diff, conf.level = 0.95)

    One Sample t-test

data:  hsb2$diff
t = -0.86731, df = 199, p-value = 0.3868
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.7841424  0.6941424
sample estimates:
mean of x
   -0.545
```

Interpreting the CI for mean difference in paired data

95% CI for the mean difference in reading and writing scores (read - write) is (-1.78, 0.69)

vs.

We are 95% confident that the average reading score is 1.78 points lower to 0.69 points higher than the average writing score.

INFERENCE FOR NUMERICAL DATA

Let's practice!

INFERENCE FOR NUMERICAL DATA

Testing for a mean with a t-test

Mine Cetinkaya-Rundel
Associate Professor of the Practice, Duke University

Hypotheses

Let's revisit the High School and Beyond survey data.

> *Do the data provide convincing evidence of a difference between the average reading and writing scores of students? Use a 5% significance level.*

$H_0 : \mu_{diff} = 0$, There is no difference between the average reading and writing scores.

$H_A : \mu_{diff} \neq 0$, There is a difference between the average reading and writing scores.

Testing for a mean with a t-test

```
t.test(hsb2$diff, null = 0, alternative = "two.sided")
```

Testing for a mean with a t-test

```
t.test(hsb2$diff, null = 0, alternative = "two.sided")

    One Sample t-test

data:  hsb2$diff
t = -0.86731, df = 199, p-value = 0.3868
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.7841424  0.6941424
sample estimates:
mean of x
   -0.545
```

INFERENCE FOR NUMERICAL DATA

Let's practice!