Mean Inference

Estimating a population average

Download the section 13 .Rmd handout to
STAT240/lecture/sect13-single-mean.

Material in this section is covered by Chapter 13 on
the notes website.

Download the file TIM.txt to STAT240/data.

The Boston Marathon is a prestigious annual 26.2 mile race.

It is held in April, but was held in October in 2011.

`TIM.txt` contains times of Boston Marathon runners from 2010 to 2011.

Possible questions of interest:

- What is the average running time μ?
- Is there a difference between μ_{2010} and μ_{2011}?

Analysis is based on the CLT.

We assume the observations, X_i are i.i.d. with mean μ and sd σ.

$$X_i \sim D(\mu, \sigma)$$

Let's focus on the 18-34 female finishers from 2010.

The point estimate for μ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and we observed $\bar{x} = 235.5$.

Let's look at a simulation for $\bar{X}$.

The CLT gives

$$\bar{X} \overset{.}{\sim} N\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right)$$

for large enough n, which will be the basis of our CI and H-test.

A CI, in general, is

point estimate $\pm$ critical value $\times$ standard error

- For μ, the point estimate is $\bar{x}$
- The standard error is $\frac{\sigma}{\sqrt{n}}$

We should be able to use a Z critical value:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \ \sim \ N(0, 1)$$

Problem: we don't have σ. The quantity

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$
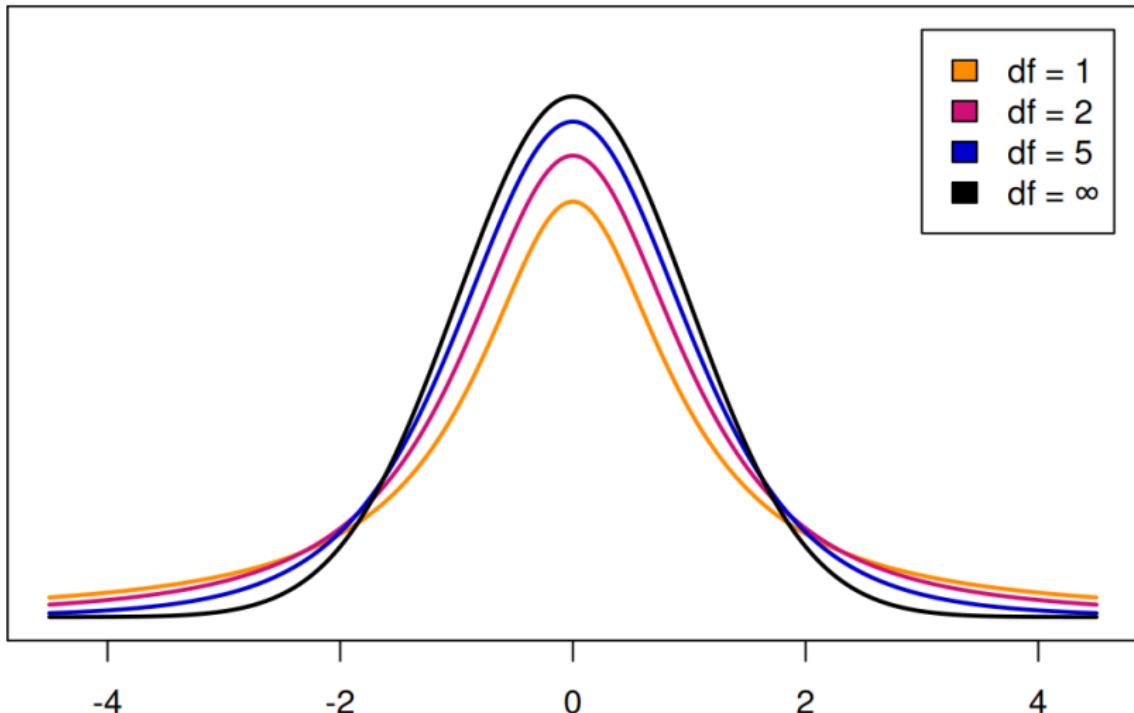
is not *quite* normal.

The sampling distribution for $\bar{X}$ when σ is unknown uses the **Student's T distribution**.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T$$

This is similar to $N(0, 1)$, but wider.

T distributions
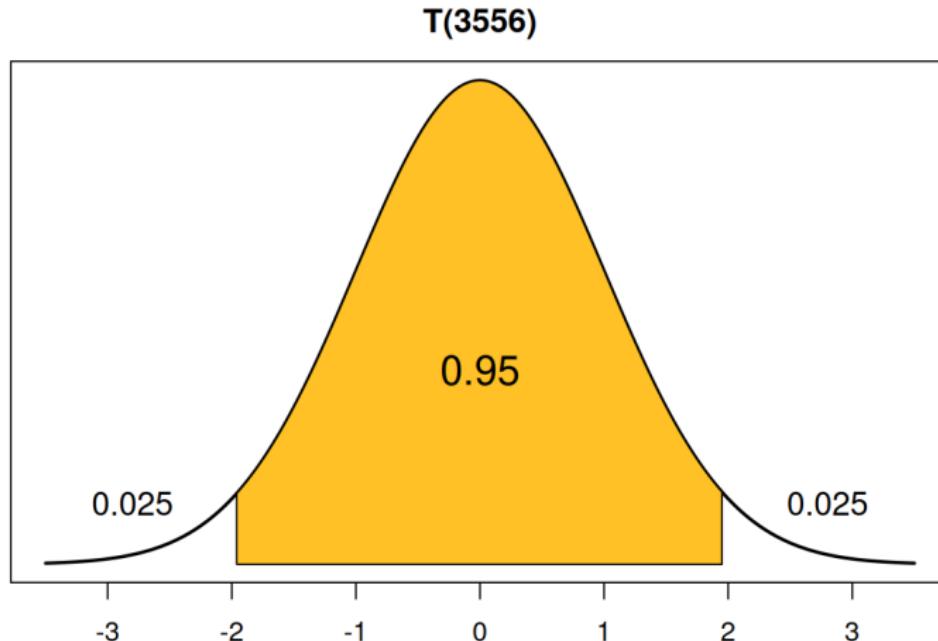
The T has heavier tails than $N(0, 1)$, controlled by degrees of freedom.

In single mean inference, $df = n - 1$.

Find critical values (quantiles) with qt().

For 95% confidence:

T(3556)

T CI for μ:

$$\bar{X} \ \pm \ t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$$

We are 95% confident that the true average time for an 18-34 year old female runner in 2010 is within (234.3, 236.7).

Check with `t.test()`.

Is the true average time for an 18-34 year old female runner in 2010 equal to 240 minutes?

$$H_0 : \mu = 240 \quad \text{versus} \quad H_A : \mu \neq 240$$

We gather evidence against H_0 with $\bar{x} = 235.5$.

Let's use $\alpha = 0.05$.

If $\mu = 240$, then the quantity

$$\frac{\bar{X} - 240}{\sigma/\sqrt{n}}$$

should be $N(0, 1)$, implying a Z test.

But, we don't know σ.
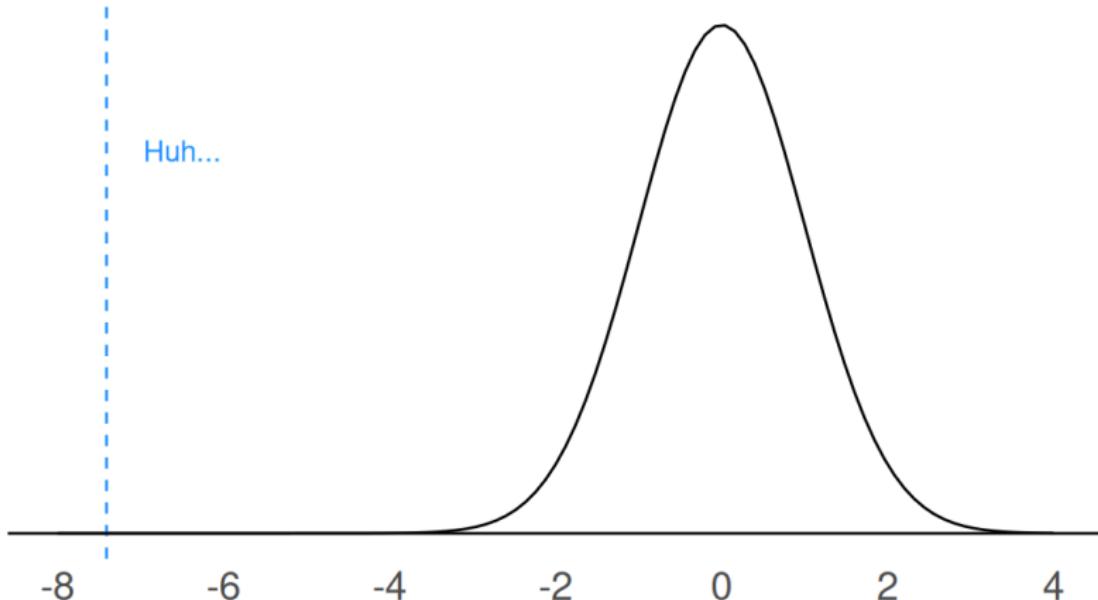
Instead, use

$$\frac{\bar{X} - 240}{S/\sqrt{n}} \sim T_{n-1}$$

Our observed test statistic is -7.41. Is this consistent with our null T distribution?

T(3556)

Huh...

-8   -6   -4   -2   0   2   4
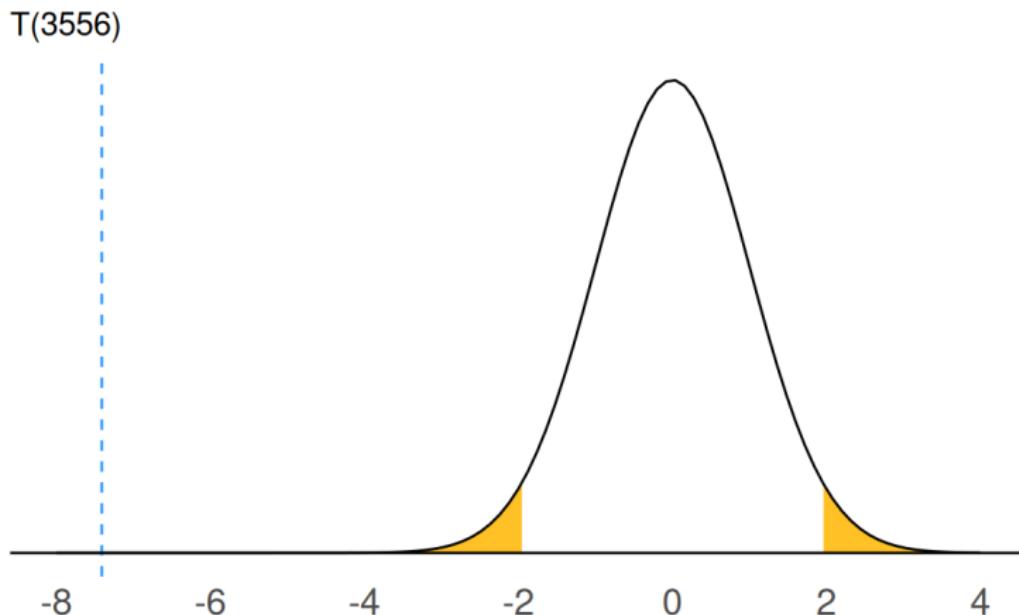
Our p-value

$$P(T_{n-1} \leq -7.41) \quad + \quad P(T_{n-1} \geq 7.41)$$

is very small, so we reject H_0.

We can check this result in `t.test()`.

A rejection region also rejects H_0.

Recall that our 95% CI for μ did not contain 240: (234.3, 236.7).

Similarly, we rejected the null value 240 at the 5% level. In fact, the methods are equivalent.

The following 2 statements are equivalent for T CIs
and hypothesis tests for μ:

- A $100(1 - \alpha)\%$ CI contains μ_0.
- A test for $H_A : \mu \neq \mu_0$ at level α rejects H_0.

Conversely, if μ_0 is inside of the $100(1 - \alpha)\%$ CI, we
would fail to reject H_0 in a two-sided test.

Consider female runners age 60-64 from 2010. Do we have evidence that their average time is *less* than 280 minutes?

- Write one-sided hypotheses about μ
- Re-calculate summary statistics
- Perform the T test with $\alpha = 0.02$.
- Check with `t.test()`

In general, perform a T test on independent draws from the same population. Use hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_A : \mu \neq \mu_0$$

and test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

and null distribution T_{n-1}.

P-value calculation:

- If the alternative is $H_A : \mu < \mu_0$, the p-value is the area *below* the test statistic.

- If the alternative is $H_A : \mu > \mu_0$, the p-value is the area *above* the test statistic.

- If the alternative is $H_A : \mu \neq \mu_0$, the p-value is $2\times$ the area *outside* of the test statistic.