Inference for a Single Mean

Estimating a population average

Download the section 16 .Rmd handout to `STAT240/lecture/sect16-single-mean`.

Download the file `TIM.txt` to `STAT240/data`.

The Boston Marathon is a prestigious 26.2 mile annual 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

- Average running time for a given year
- Change in running time from 2010 to 2011

Let's tidy and explore the data first.

What is the average time μ of all finishers? Write each individual observation as a random variable X_i.

We must assume that each X_i is independent with the same distribution with mean μ and sd σ.

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

Let's focus on the 3557 18-34 female finishers from 2010. The point estimate for μ is the sample mean

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

By the CLT, this variable is approximately normal for large enough n.

Our observed mean is $\bar{x} = 235.5$. How reliable is this guess?

The theoretical mean $\bar{X}$ has $E(\bar{X} = \mu$, $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. The CLT gives

$$\bar{X} \dot{\sim} N\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big)$$

which will be the basis of our CI and H-test.

A CI, in general, is

$$\text{point estimate} \pm \text{critical value} \times \text{standard error}$$

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

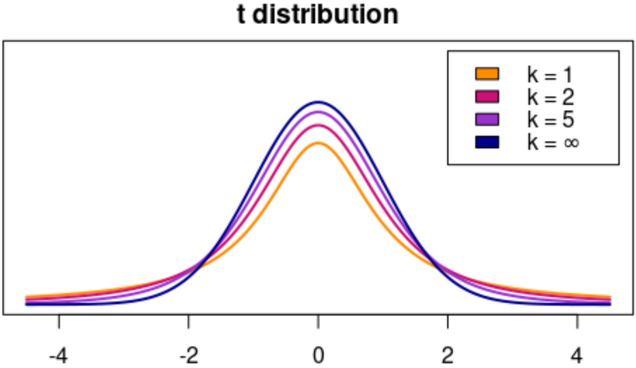
We should be able to use a Z critical value:

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

Problem: we don't have σ. Instead, use a T critical value.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \;\dot{\sim}\; t_{n-1}$$

The T bell curve is wider than $N(0, 1)$.



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 need to gather evidence against H_0 with our observed $\bar{x} = 235.5$.

If $\mu = 240$, then $\bar{X}$ should be close to 240, and $\bar{X} - 240$ should be small.

We also have to take estimation error (standard error) into account.

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

should be close to 0.

Formally, our model is

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

and our null is $H_0 : \mu = 240$. The test statistic must have a known distribution if H_0 is true.

$$\frac{\bar{X} - 240}{\sigma/\sqrt{n}} \ \dot{\sim} \ N(0, 1)$$

will not work.

Instead, use

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

Our observed test statistic is -7.41. For a two-sided p-value, calculate

$$P(t_{n-1} < -7.41) + P(t_{n-1} > 7.41)$$

We get a very small p-value, so we reject H_0.

We can verify this result in `t.test` by specifying our null value.

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.