STA286 Lecture 30

Neil Montgomery

Last edited: 2017-04-03 13:57

hypothesis testing

context: a one-sample t interval example

Let's suppose it is widely known ("everyone knows") that the healthy amount of `Fe` to have in engine oil is, say, "4ppm".

You have $n = 30$ haul trucks and you want to assess the health of the fleet of engines.

context: a one-sample t interval example

Let's suppose it is widely known ("everyone knows") that the healthy amount of `Fe` to have in engine oil is, say, "4ppm".

You have $n = 30$ haul trucks and you want to assess the health of the fleet of engines.

You take the oil samples and you end up with a sample average of $\overline{x} = 7.15$ ppm and a sample standard deviation of $s = 8.82$ ppm.

context: a one-sample t interval example

Let's suppose it is widely known ("everyone knows") that the healthy amount of `Fe` to have in engine oil is, say, "4ppm".

You have $n = 30$ haul trucks and you want to assess the health of the fleet of engines.

You take the oil samples and you end up with a sample average of $\overline{x} = 7.15$ ppm and a sample standard deviation of $s = 8.82$ ppm.

A 95% confidence interval for the mean amount of `Fe` in your fleet of engines is easily computed to be:

$$7.15 \pm t_{29,0.025} \frac{8.82}{\sqrt{30}}$$

or [3.86, 10.44].

context: a one-sample t interval example

Let's suppose it is widely known ("everyone knows") that the healthy amount of `Fe` to have in engine oil is, say, "4ppm".

You have $n = 30$ haul trucks and you want to assess the health of the fleet of engines.

You take the oil samples and you end up with a sample average of $\overline{x} = 7.15$ ppm and a sample standard deviation of $s = 8.82$ ppm.

A 95% confidence interval for the mean amount of `Fe` in your fleet of engines is easily computed to be:

$$7.15 \pm t_{29,0.025} \frac{8.82}{\sqrt{30}}$$

or [3.86, 10.44].

Your engines are not healthy!

context: a one-sample t interval example

Let's suppose it is widely known ("everyone knows") that the healthy amount of `Fe` to have in engine oil is, say, "4ppm".

You have $n = 30$ haul trucks and you want to assess the health of the fleet of engines.

You take the oil samples and you end up with a sample average of $\overline{x} = 7.15$ ppm and a sample standard deviation of $s = 8.82$ ppm.

A 95% confidence interval for the mean amount of `Fe` in your fleet of engines is easily computed to be:

$$7.15 \pm t_{29,0.025} \frac{8.82}{\sqrt{30}}$$

or [3.86, 10.44].

Your engines are not healthy!

"Hypothesis testing" generally involves using data to make a principled statement about a parameter value.

hypotheses

Specific statements about parameter values can have practical meanings.

In the `Fe` example, the model used for the population was $X \sim N(\mu, \sigma)$.

Some statements include:

$$\mu = 4 \qquad \mu \neq 4 \qquad \sigma = 1 \qquad \sigma^2 > 5$$

hypotheses

Specific statements about parameter values can have practical meanings.

In the `Fe` example, the model used for the population was $X \sim N(\mu, \sigma)$.

Some statements include:

$$\mu = 4 \qquad \mu \neq 4 \qquad \sigma = 1 \qquad \sigma^2 > 5$$

To be honest, when only considering one population, this all can look mysterious and arbitrary.

hypotheses

Specific statements about parameter values can have practical meanings.

In the `Fe` example, the model used for the population was $X \sim N(\mu, \sigma)$.

Some statements include:

$$\mu = 4 \qquad \mu \neq 4 \qquad \sigma = 1 \qquad \sigma^2 > 5$$

To be honest, when only considering one population, this all can look mysterious and arbitrary.

A statement about a parameter value is called a *hypothesis*.

some more natural hypotheses

Consider two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$. The most obviously interesting hypothesis is:

$$\mu_1 = \mu_2$$

which is the hypothesis that encapsulates "no difference".

some more natural hypotheses

Consider two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$. The most obviously interesting hypothesis is:

$$\mu_1 = \mu_2$$

which is the hypothesis that encapsulates "no difference".

Similarly, consider two population Bernoulli(p_1) and Bernoulli(p_2). We might also have:

$$p_1 = p_2$$

to mean "no difference"

null hypothesis and alternative hypothesis

In hypothesis testing we settle on two “hypotheses” concerning parameter values.

The only mathematical requirement is that they don’t contain any parameter values in common.

But in practice it is not so arbitrary.

null hypothesis and alternative hypothesis

In hypothesis testing we settle on two "hypotheses" concerning parameter values.

The only mathematical requirement is that they don't contain any parameter values in common.

But in practice it is not so arbitrary.

The null hypothesis, denoted by H_0, is almost always the "no effect"/"no difference"/"status quo" parameter value.

null hypothesis and alternative hypothesis

In hypothesis testing we settle on two "hypotheses" concerning parameter values.

The only mathematical requirement is that they don't contain any parameter values in common.

But in practice it is not so arbitrary.

The null hypothesis, denoted by H_0, is almost always the "no effect"/"no difference"/"status quo" parameter value.

For example, $\mu_1 = \mu_2$ and $p_1 = p_2$.

null hypothesis and alternative hypothesis

In hypothesis testing we settle on two "hypotheses" concerning parameter values.

The only mathematical requirement is that they don't contain any parameter values in common.

But in practice it is not so arbitrary.

The null hypothesis, denoted by H_0, is almost always the "no effect"/"no difference"/"status quo" parameter value.

For example, $\mu_1 = \mu_2$ and $p_1 = p_2$.

The alternative hypothesis, denoted by H_1, is usually the complement of H_0.

null hypothesis and alternative hypothesis

In hypothesis testing we settle on two "hypotheses" concerning parameter values.

The only mathematical requirement is that they don't contain any parameter values in common.

But in practice it is not so arbitrary.

The null hypothesis, denoted by H_0, is almost always the "no effect"/"no difference"/"status quo" parameter value.

For example, $\mu_1 = \mu_2$ and $p_1 = p_2$.

The alternative hypothesis, denoted by H_1, is usually the complement of H_0.

For example, $\mu_1 \neq \mu_2$ and $p_1 \neq p_2$.

opinion - The Myth of the "One-Sided" Alternative

Textbooks go on about "choosing" the "appropriate" alternative hypothesis, based on little more than the hopes and dreams of the experimenter.

opinion - The Myth of the “One-Sided” Alternative

Textbooks go on about “choosing” the “appropriate” alternative hypothesis, based on little more than the hopes and dreams of the experimenter.

Students are sent on wild good chases trying to guess what the textbook author/instructor is “hoping” the alternative is.

In my **opinion** this is nonsense. The alternative should almost always be the completment of the null.

opinion - The Myth of the “One-Sided” Alternative

Textbooks go on about “choosing” the “appropriate” alternative hypothesis, based on little more than the hopes and dreams of the experimenter.

Students are sent on wild good chases trying to guess what the textbook author/instructor is “hoping” the alternative is.

In my **opinion** this is nonsense. The alternative should almost always be the completment of the null.

(Note: this is a scientific opinion, and not a mathematical opinion.)

“classical” hypothesis testing

Our first view of hypothesis testing has a clear goal, which is to use data to make a specific decision: to either *reject* H_0 or *not reject* H_0.

Sometimes *accept* is used as a synonym for *not reject*. The book uses the phrase *fail to reject*, which I’ve never seen anywhere alse.

The main thing is to avoid attaching positive or negative connotations to any of these phrases.

“classical” hypothesis testing

Our first view of hypothesis testing has a clear goal, which is to use data to make a specific decision: to either *reject* H_0 or *not reject* H_0.

Sometimes *accept* is used as a synonym for *not reject*. The book uses the phrase *fail to reject*, which I've never seen anywhere alse.

The main thing is to avoid attaching positive or negative connotations to any of these phrases.

The method in a nutshell: assume H_0, collect a sample, and see if the sample contradicts H_0.

"classical" hypothesis testing

Our first view of hypothesis testing has a clear goal, which is to use data to make a specific decision: to either *reject* H_0 or *not reject* H_0.

Sometimes *accept* is used as a synonym for *not reject*. The book uses the phrase *fail to reject*, which I've never seen anywhere alse.

The main thing is to avoid attaching positive or negative connotations to any of these phrases.

The method in a nutshell: assume H_0, collect a sample, and see if the sample contradicts H_0.

Motivating example... a pre-fabricated furniture company needs its supplier to provide doors that are 700mm wide. Does the supplier meet this target?

"classical" hypothesis testing

Our first view of hypothesis testing has a clear goal, which is to use data to make a specific decision: to either *reject* H_0 or *not reject* H_0.

Sometimes *accept* is used as a synonym for *not reject*. The book uses the phrase *fail to reject*, which I've never seen anywhere alse.

The main thing is to avoid attaching positive or negative connotations to any of these phrases.

The method in a nutshell: assume H_0, collect a sample, and see if the sample contradicts H_0.

Motivating example... a pre-fabricated furniture company needs its supplier to provide doors that are 700mm wide. Does the supplier meet this target?

The model for door width will be $N(\mu, \sigma)$, with $\sigma = 0.5$ magically known for now.

classical hypothesis testing—motivating example

The null and alternative hypotheses are:

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

We plan to gather a sample of size $n = 10$.

classical hypothesis testing—motivating example

The null and alternative hypotheses are:

$$
\begin{aligned}
H_0 &: \mu = 700 \\
H_1 &: \mu \neq 700
\end{aligned}
$$

We plan to gather a sample of size $n = 10$.

What *statistic* should be used to make statements about μ. Probably a good idea to use the MLE $\overline{X}$. This is called the *test statistic*.

classical hypothesis testing—motivating example

The null and alternative hypotheses are:

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

We plan to gather a sample of size $n = 10$.

What *statistic* should be used to make statements about μ. Probably a good idea to use the MLE $\overline{X}$. This is called the *test statistic*.

Suppose (temporarily, as a thought experiment) that in fact $\mu = 700$. What is the distrubtion of $\overline{X}$ and which values of $\overline{X}$ would surprise us?

classical hypothesis testing—motivating example

The null and alternative hypotheses are:

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

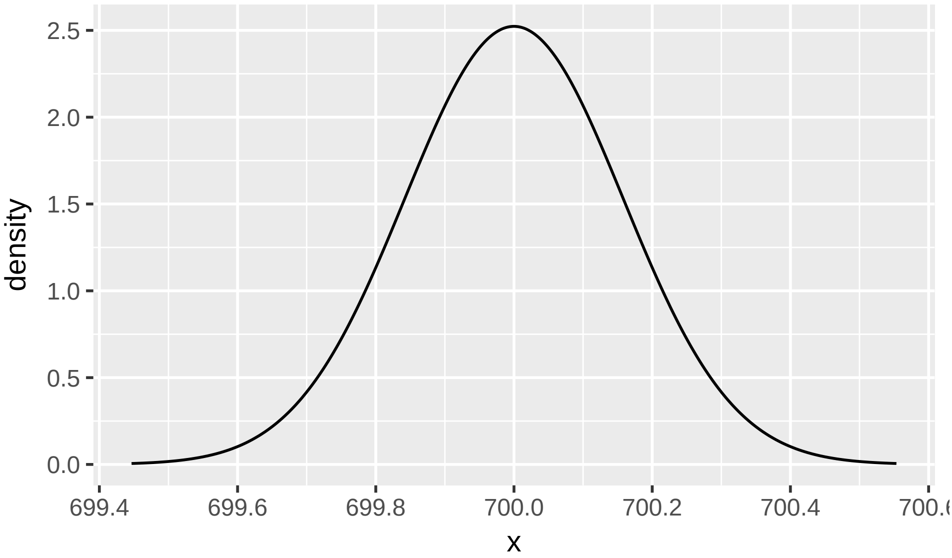
We plan to gather a sample of size $n = 10$.

What *statistic* should be used to make statements about μ. Probably a good idea to use the MLE $\overline{X}$. This is called the *test statistic*.

Suppose (temporarily, as a thought experiment) that in fact $\mu = 700$. What is the distrubtion of $\overline{X}$ and which values of $\overline{X}$ would surprise us?

$$\overline{X} \sim N\left(700, \frac{0.5}{\sqrt{10}}\right) \quad \text{The } \textit{null distribution}$$

null distribution $N(700, 0.1581139)$



classical hypothesis testing - the details

"The values that would surprise us" are defined in advance according to a pre-set probabilty α.

This is called the "size" of the test, or the "level of significance". It is typically something small like: 0.05, 0.1, 0.05, 0.05, 0.01, 0.05, or 0.05.

classical hypothesis testing - the details

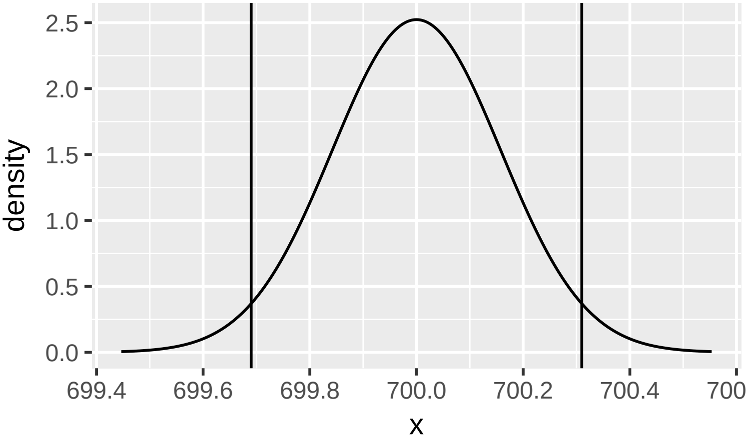
"The values that would surprise us" are defined in advance according to a pre-set probabilty α.

This is called the "size" of the test, or the "level of significance". It is typically something small like: 0.05, 0.1, 0.05, 0.05, 0.01, 0.05, or 0.05.

α is the *probability of rejecting* H_0 *when it is in fact true.*

classical hypothesis testing - the details

Suppose $\alpha = 0.05$. The "area of surprise" in our motivating example is defined as $\overline{X} \leqslant 699.6901$ or $\overline{X} \geqslant 700.3099$, as in:



classical hypothesis testing - the details

The "area of surprise" is really called the "rejection region" or "critical region".

We define two types of "error" in classical hypothesis testing:

	Action	
"Truth"	Reject	Not Reject
H_0 True		
H_0 False		

classical hypothesis testing - the details

The "area of surprise" is really called the "rejection region" or "critical region".

We define two types of "error" in classical hypothesis testing:

	Action	
"Truth"	Reject	Not Reject
H_0 True	Type I Error	
H_0 False		

classical hypothesis testing - the details

The “area of surprise” is really called the “rejection region” or “critical region”.

We define two types of “error” in classical hypothesis testing:

	Action	
“Truth”	Reject	Not Reject
H_0 True	Type I Error	
H_0 False		Type II Error

classical hypothesis testing - the details

The "area of surprise" is really called the "rejection region" or "critical region".

We define two types of "error" in classical hypothesis testing:

	Action	
"Truth"	Reject	Not Reject
H_0 True	Type I Error	
H_0 False		Type II Error

$$\alpha = P(\text{Type I Error}) \qquad \beta = P(\text{Type II Error})$$

classical hypothesis testing - the details

The "area of surprise" is really called the "rejection region" or "critical region".

We define two types of "error" in classical hypothesis testing:

	Action	
"Truth"	Reject	Not Reject
H_0 True	Type I Error	
H_0 False		Type II Error

$$\alpha = P(\text{Type I Error}) \qquad \beta = P(\text{Type II Error})$$

The probability $1 - \beta$ of rejecting H_0 when it is false is called the "power" of the test.

example of critical region

In our motivating example, the critical region comes from this expression that uses the null distribution:

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - 700}{0.5/\sqrt{10}} < z_{\alpha/2}\right) = 1 - \alpha$$

example of critical region

In our motivating example, the critical region comes from this expression that uses the null distribution:

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - 700}{0.5/\sqrt{10}} < z_{\alpha/2}\right) = 1 - \alpha$$

The region is:

$$\left\{\overline{X} < 700 - z_{\alpha/2}\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + z_{\alpha/2}\frac{0.5}{\sqrt{10}}\right\}$$

example of critical region

In our motivating example, the critical region comes from this expression that uses the null distribution:

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - 700}{0.5/\sqrt{10}} < z_{\alpha/2}\right) = 1 - \alpha$$

The region is:

$$\left\{\overline{X} < 700 - z_{\alpha/2}\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + z_{\alpha/2}\frac{0.5}{\sqrt{10}}\right\}$$

If we set $\alpha = 0.01$, say, this becomes:

$$\left\{\overline{X} < 699.593\right\} \cup \left\{\overline{X} > 700.407\right\}$$

example power calculation

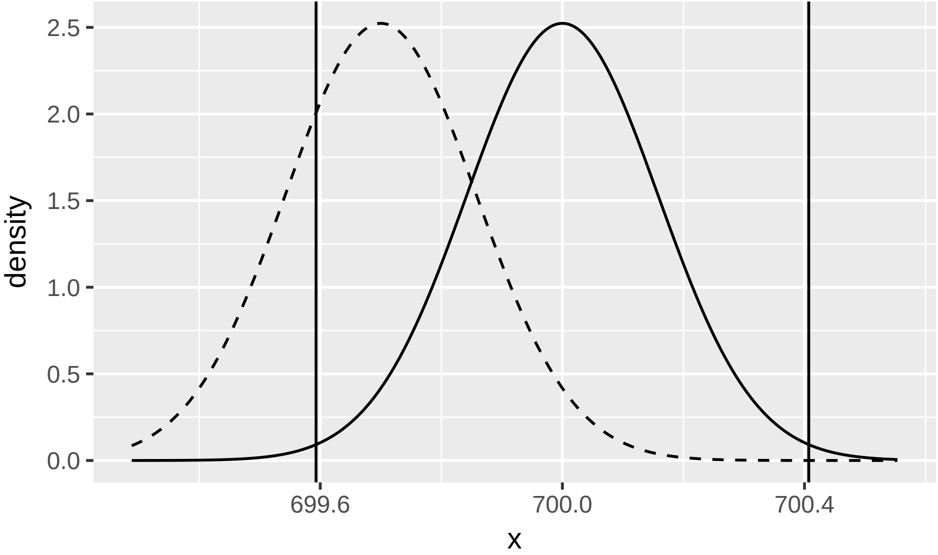
For an explicit power calculation, one needs a specific alternative.

So, suppose in fact the supplier makes doors that are $\mu_1 = 699.7$mm wide. So in fact $\overline{X} \sim N(699.7, 0.5/\sqrt(10))$

What is the probability of "rejecting H_0"?

$$\begin{aligned} P_{\mu_1}(\overline{X} < 699.593) + P_{\mu_1}(\overline{X} > 700.407) &= P(Z < -0.677) + P(Z > 4.471) \\ &= 0.249 + 0 \end{aligned}$$

power in pictures



size, power, and sample size

When the population is $N(\mu, \sigma)$ and the sample is $X_1, \ldots, X_n$ and the hypotheses are $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, the generic rejection region is, for fixed α:

$$\left\{\overline{X} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} \cup \left\{\overline{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$