Tests of significance:

The basics

BPS chapter 15

Objectives (BPS chapter 15)

Tests of significance: the basics

- The reasoning of tests of significance
- Stating hypotheses
- Test statistics
- *P*-values
- Statistical significance
- Tests for a population mean
- Using tables of critical values
- Tests from confidence intervals

We have seen that the properties of the sampling distribution of *x* bar help us estimate a range of likely values for population mean μ.

We can also rely on the properties of the sample distribution to test hypotheses.

Example: You are in charge of quality control in your food company. You sample randomly four packs of cherry tomatoes, each labeled 1/2 lb. (227 g).

The average weight from your four boxes is 222 g. Obviously, we cannot expect boxes filled with whole tomatoes to all weigh exactly half a pound. Thus:

- Is the somewhat smaller weight simply due to chance variation?
- Is it evidence that the calibrating machine that sorts cherry tomatoes into packs needs revision?

Hypotheses tests

A **test of statistical significance** tests a specific hypothesis using sample data to decide on the validity of the hypothesis.

In statistics, a **hypothesis** is an assumption, or a theory about the characteristics of one or more variables in one or more populations.

What you want to know: Does the calibrating machine that sorts cherry tomatoes into packs need revision?

The same question reframed statistically: Is the population mean μ for the distribution of weights of cherry tomato packages equal to 227 g (i.e., half a pound)?

The **null hypothesis** is the statement being tested. It is a statement of "no effect" or "no difference," and it is labeled H_0.

The **alternative hypothesis** is the claim we are trying to find evidence ***for***, and it is labeled H_a.

Weight of cherry tomato packs:

H_0: μ = 227 g (μ is the average weight of the population of packs)

H_a: $\mu \neq$ 227 g (μ is either larger or smaller)

One-sided and two-sided tests

- A **two-tail** or **two-sided test** of the population mean has these null and alternative hypotheses:

H_0: μ = [a specific number] H_a: $\mu \neq$ [a specific number]

- A **one-tail** or **one-sided test** of a population mean has these null and alternative hypotheses:

H_0: μ = [a specific number] H_a: $\mu <$ [a specific number] OR

H_0: μ = [a specific number] H_a: $\mu >$ [a specific number]

The FDA tests whether a generic drug has an absorption extent similar to the known absorption extent of the brand-name drug it is copying. Higher or lower absorption would both be problematic, thus we test:

H_0: $\mu_{generic} = \mu_{brand}$ H_a: $\mu_{generic} \neq \mu_{brand}$ two-sided

How to choose?

What determines the choice of a one-sided versus two-sided test is what we know about the problem before we perform a test of statistical significance.

A health advocacy group tests whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 mg.

Here, the health advocacy group suspects that cigarette manufacturers sell cigarettes with a nicotine content higher than what they advertise in order to better addict consumers to their products and maintain revenues.

Thus, this is a one-sided test: H_0: μ = 1.4 mg H_a: μ > 1.4 mg

It is important to make that choice before performing the test or else you could make a choice of "convenience" or fall in circular logic.

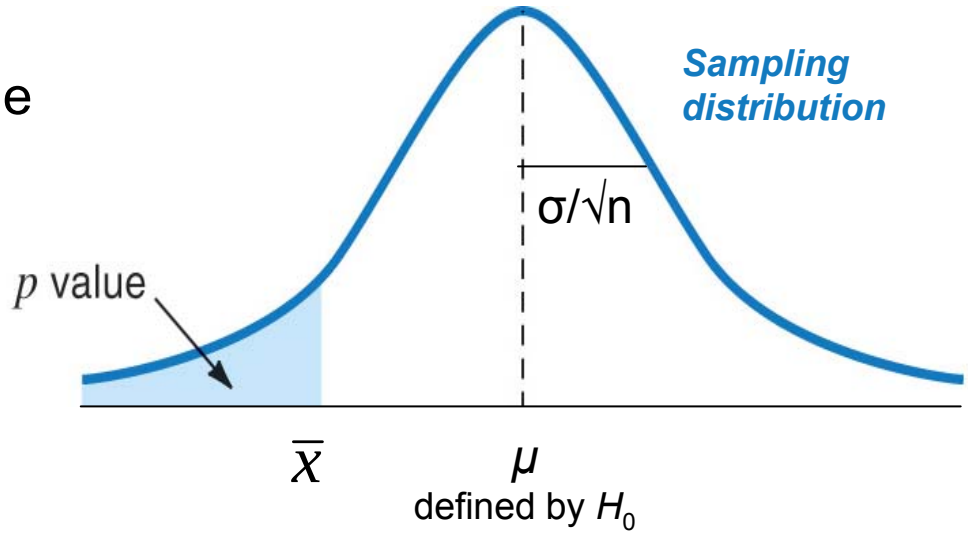
Tests for a population mean

To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n from a Normal population with unknown mean μ and known standard deviation σ, we rely on the properties of the sampling distribution $N(\mu, \sigma\sqrt{n})$.

The *P*-value is the area under the sampling distribution for values at least as extreme, in the direction of H_a, as that of our random sample.

Again, we first calculate a *z*-value and then use Table A.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



The *P*-value

The packaging process has a known standard deviation $\sigma = 5$ g.

H_0: $\mu = 227$ g versus H_a: $\mu \neq 227$ g

The average weight from your four random boxes is 222 g.

What is the probability of drawing a random sample such as yours if H_0 is true?

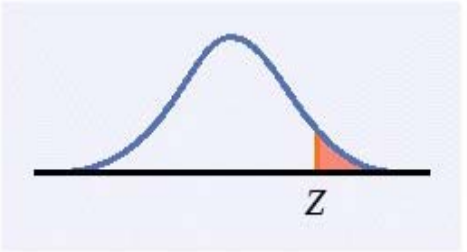
Tests of statistical significance quantify the chance of obtaining a particular random sample result if the null hypothesis were true. This quantity is the ***P*-value**.

This is a way of assessing the "believability" of the null hypothesis given the evidence provided by a random sample.

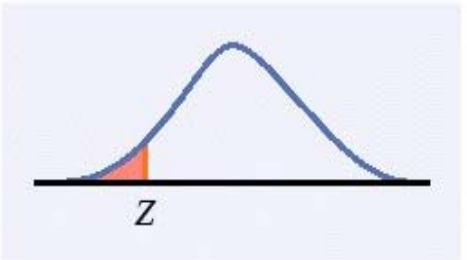
P-value in one-sided and two-sided tests

One-sided (one-tailed) test

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

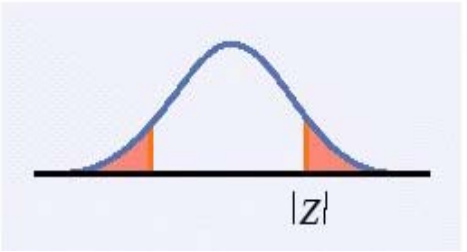


$H_a: \mu < \mu_0$ is $P(Z \leq z)$



Two-sided (two-tailed) test

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



To calculate the *P*-value for a two-sided test, use the symmetry of the normal curve. Find the *P*-value for a one-sided test and double it.

Interpreting a *P*-value

Could random variation alone account for the difference between the null hypothesis and observations from a random sample?

- A small *P*-value implies that random variation because of the sampling process alone is not likely to account for the observed difference.

- With a small *P*-value, we **reject H_0**. The true property of the population is **significantly** different from what was stated in H_0.

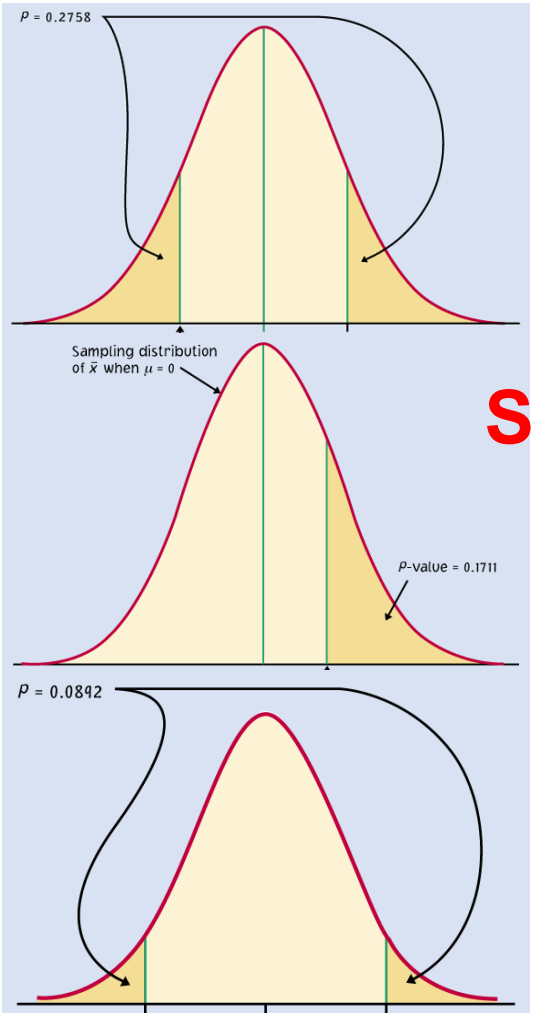
Thus small *P*-values are strong evidence AGAINST H_0.

But how small is small…?

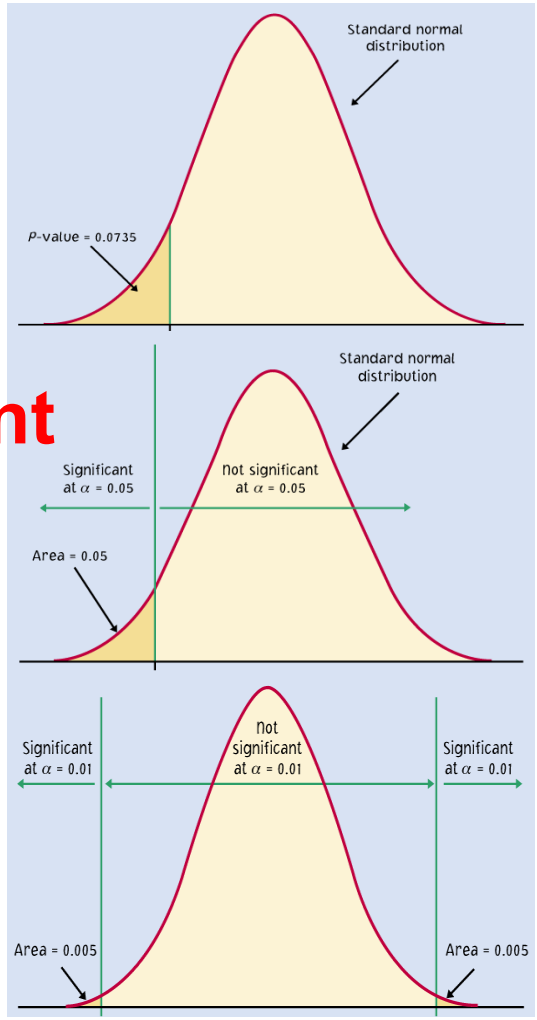
P = 0.2758

P = 0.1711

P = 0.0892



Significant *P*-value ???



P = 0.0735

P = 0.05

P = 0.01

When the shaded area becomes very small, the probability of drawing such a sample at random gets very slim. Oftentimes, a *P*-value of 0.05 or less is considered **significant:** The phenomenon observed is unlikely to be entirely due to chance event from the random sampling.

Does the packaging machine need revision?

- H_0: μ = 227 g versus H_a: $\mu \neq 227$ g
- What is the probability of drawing a random sample such as yours if H_0 is true?

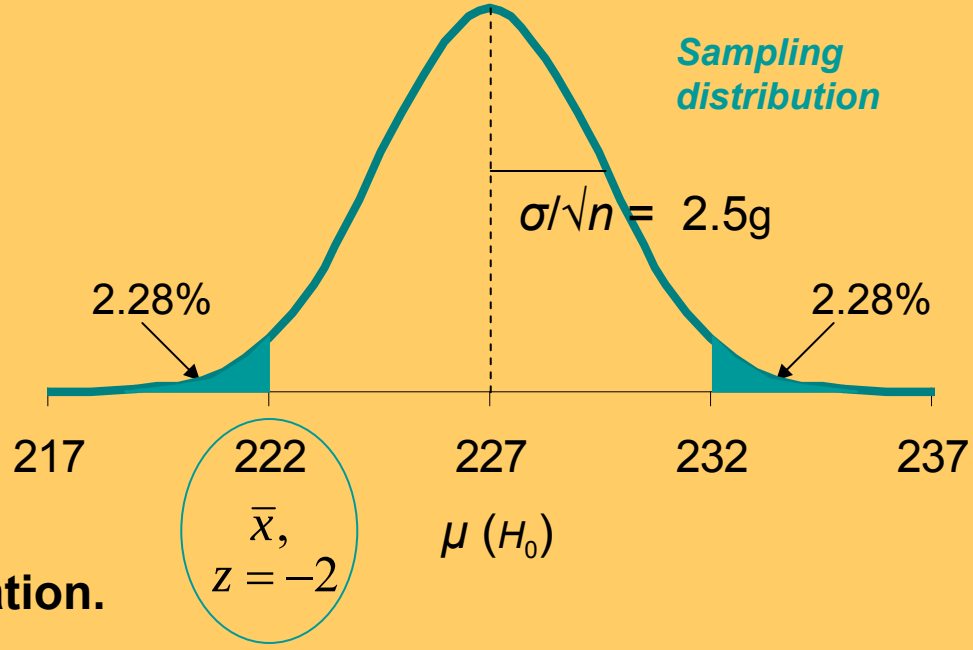
$$\bar{x} = 222\text{g} \quad \sigma = 5\text{g} \quad n = 4 \qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{222 - 227}{5/\sqrt{4}} = -2$$

From Table A, the area under the standard normal curve to the left of *z* is 0.0228.

Thus, *P*-value = 2*0.0228 = 4.56%.

The probability of getting a random sample average so different from μ is so low that we reject H_0.

➔**The machine does need recalibration.**



The significance level α

The significance level, α, is the largest *P*-value tolerated for rejecting a true null hypothesis (how much evidence against H_0 we require). This value is decided arbitrarily before conducting the test.

- If the *P*-value is equal to or less than α ($p \leq \alpha$), then we **reject H_0**.
- If the *P*-value is greater than α ($p > \alpha$), then we **fail to reject H_0**.

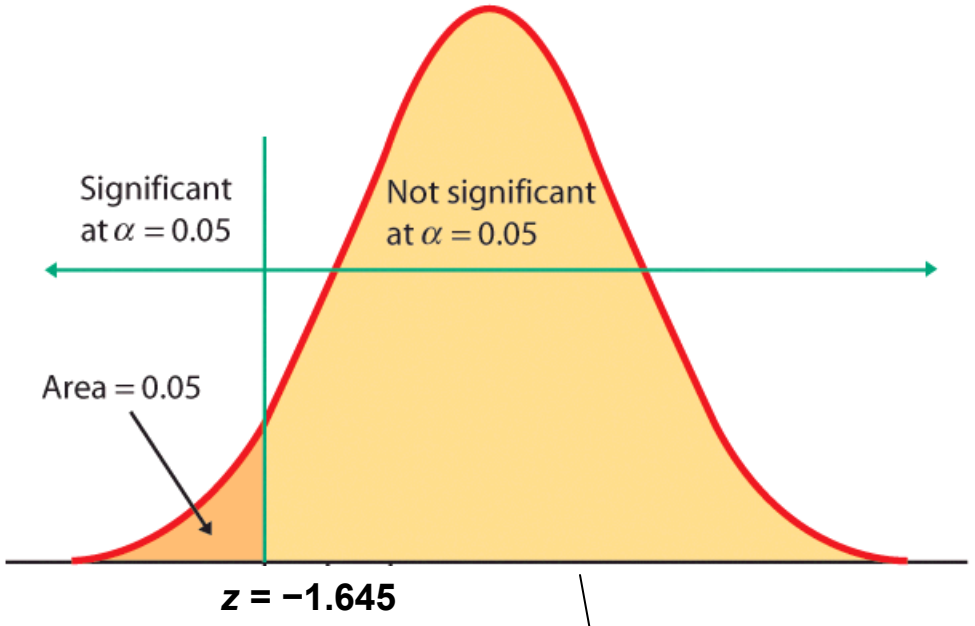
Does the packaging machine need revision?

Two-sided test. The *P*-value is 4.56%.

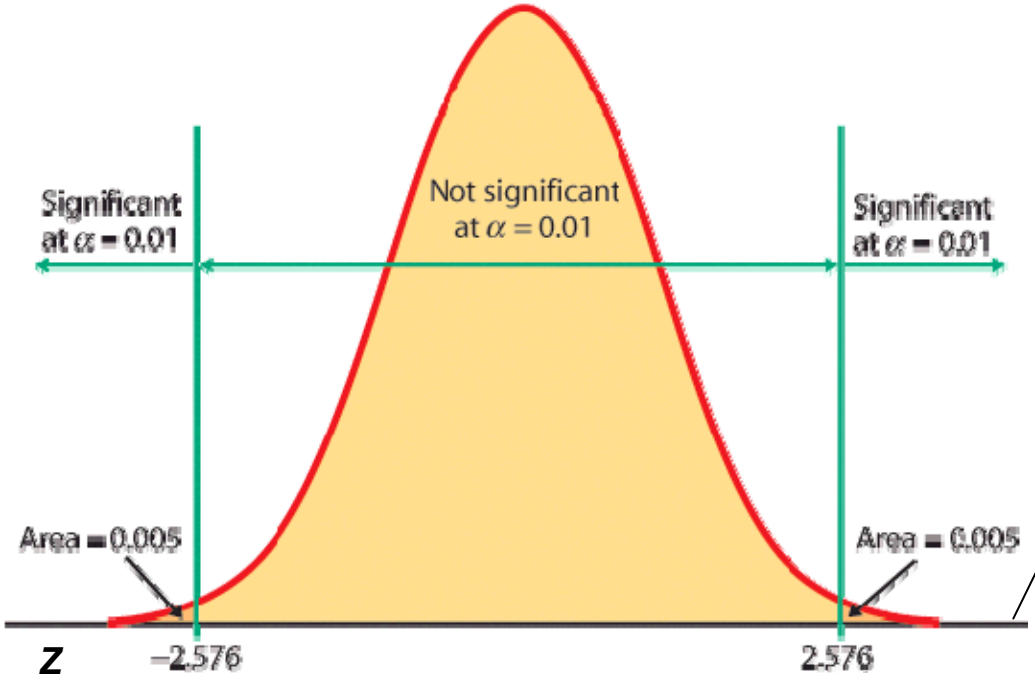
* If α had been set to 5%, then the *P*-value would be significant.

* If α had been set to 1%, then the *P*-value would not be significant.

When the *z* score falls within the rejection region (shaded area on the tail-side), the *P*-value is smaller than α and you have shown statistical significance.



One-sided test, α = 5%



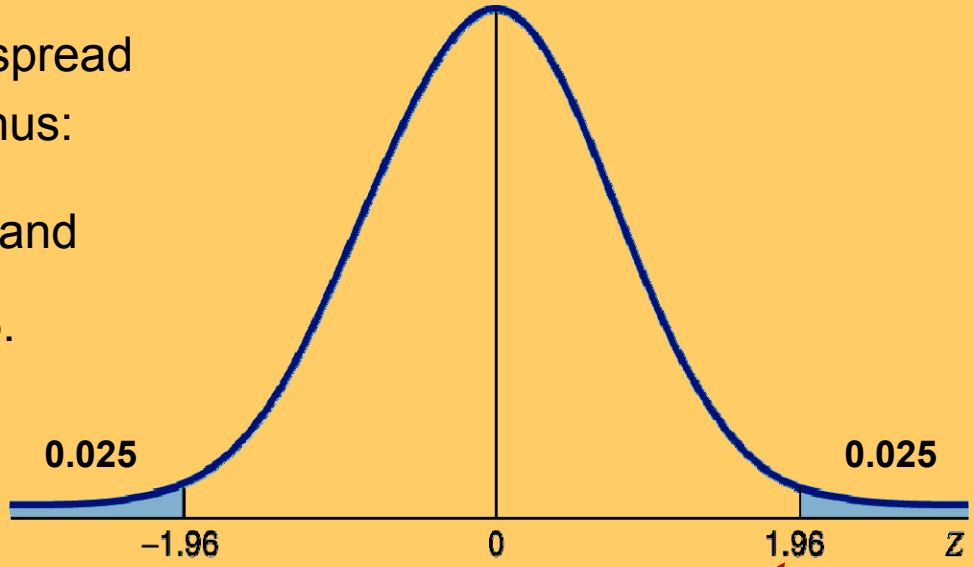
Two-sided test, α = 1%

Rejection region for a two-tail test of μ with α = 0.05 (5%)

A two-sided test means that α is spread between both tails of the curve, thus:

-a middle area *C* of 1 − *α* = 95%, and

-an upper tail area of *α* /2 = 0.025.



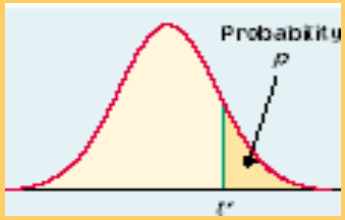


Table C

upper tail probability	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
(…)												
z*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
Confidence interval C	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Confidence intervals to test hypotheses

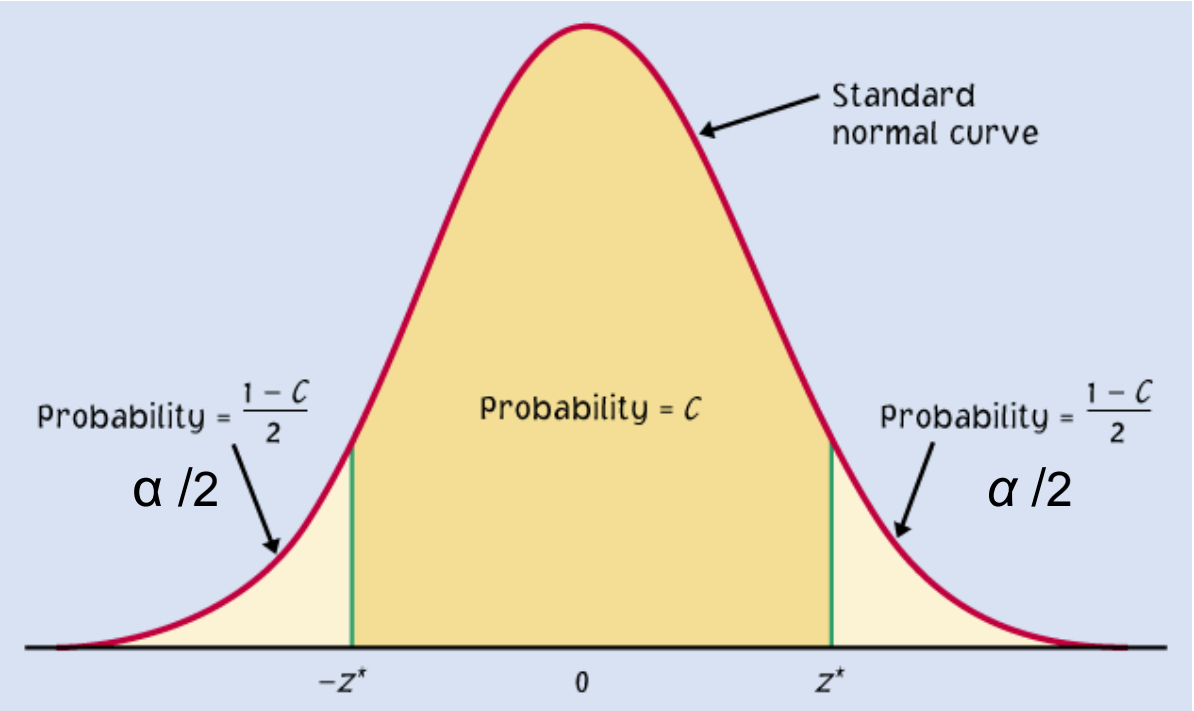
Because a two-sided test is symmetrical, you can also use a confidence interval to test a two-sided hypothesis.

In a two-sided test,

$C = 1 - \alpha$.

C confidence level

α significance level



Packs of cherry tomatoes (σ = 5 g): H_0: μ = 227 g versus H_a: $\mu \neq 227$ g

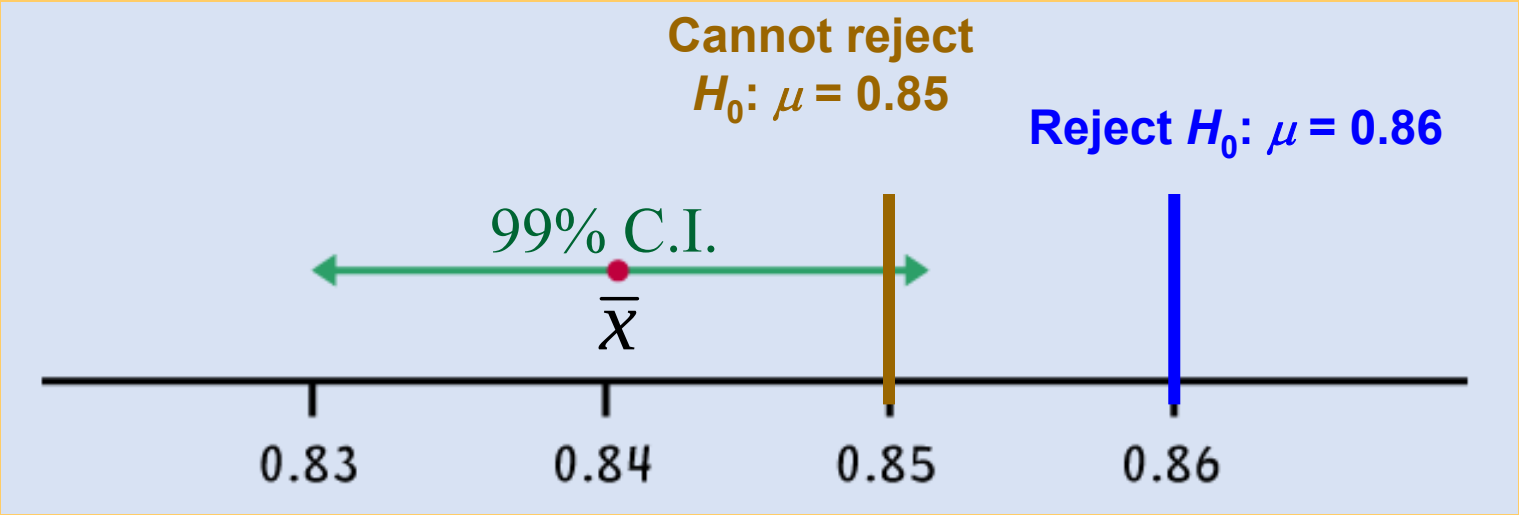
Sample average 222 g. 95% CI for $\mu = 222 \pm 1.96*5/\sqrt{4} = 222 \text{ g} \pm 4.9$ g

227 g does not belong to the 95% CI (217.1 to 226.9 g). Thus, we reject H_0.

Logic of confidence interval test

Ex: Your sample gives a 99% confidence interval of $\bar{x} \pm m = 0.84 \pm 0.0101$.

With 99% confidence, could samples be from populations with μ =0.86? μ =0.85?



A confidence interval gives a black and white answer: Reject or don't reject H_0.

But it also estimates a range of likely values for the true population mean μ.

A *P*-value quantifies how strong the evidence is against the H_0. But if you reject H_0, it doesn't provide any information about the true population mean μ.