Chapter 5: Hypothesis testing

- Hypothesis testing is about making decisions
- Is a hypothesis true or false?
- Are women paid less, on average, than men?

Principles of hypothesis testing

- The null hypothesis is initially *presumed* to be true
- Evidence is gathered, to see if it is consistent with the hypothesis
- If it is, the null hypothesis continues to be considered 'true' (later evidence might change this)
- If not, the null is rejected in favour of the alternative hypothesis

Two possible types of error

- Decision making is never perfect and mistakes can be made
  - Type I error: rejecting the null when true
  - Type II error: accepting the null when false

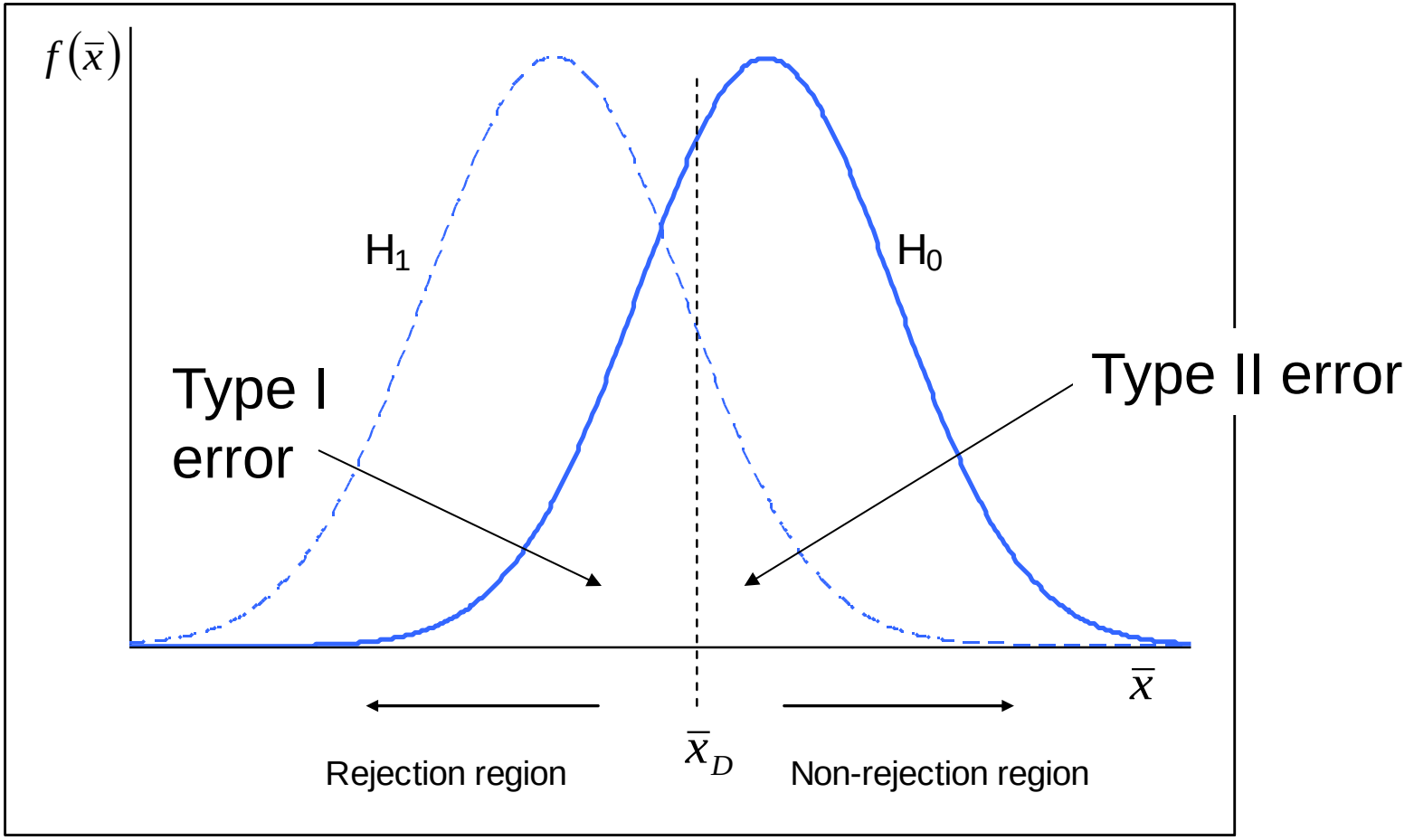
Type I and Type II errors

	True situation	
Decision	H_0 true	H_0 false
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Avoiding incorrect decisions

- We wish to avoid both Type I and II errors
- We can alter the decision rule to do this
- Unfortunately, reducing the chance of making a Type I error generally means increasing the chance of a Type II error
- Hence a trade off

Diagram of the decision rule



How to make a decision

- Where do we place the decision line?
- Set the Type I error probability to a particular value. By convention, this is 5%.
- This is known as the significance level of the test. It is complementary to the confidence level of estimation.
- 5% significance level ≡ 95% confidence level.

Example: How long do LEDs last?

- A manufacturer of LEDs claims its product lasts at least 5,000 hours, on average.
- A sample of 50 LEDs is tested. The average time before failure is 4,900 hours, with standard deviation 500 hours.
- Should the manufacturer's claim be accepted or rejected?

The hypotheses to be tested

- H_0: $\mu = 5{,}000$
  H_1: $\mu < 5{,}000$

- This is a one tailed test, since the rejection region occupies only one side of the distribution

Should the null hypothesis be rejected?

- Is 4,900 far enough below 5,000?
- Is it more than 1.64 standard errors below 5,000? (1.64 standard errors below the mean cuts off the bottom 5% of the Normal distribution.)

$$z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} = \frac{4{,}900 - 5{,}000}{\sqrt{500^2/80}} = -1.79$$

Should the null hypothesis be rejected? (continued)

- 4,900 is 1.79 standard errors below 5,000, so falls into the rejection region (bottom 5% of the distribution)
- Hence, we can reject H_0 at the 5% significance level or, equivalently, with 95% confidence.
- *If* the true mean were 5,000, there is less than a 5% chance of obtaining sample evidence such as $\bar{x} = 4{,}900$ from a sample of $n = 80$.

Formal layout of a problem

1. H_0: $\mu = 5{,}000$
   H_1: $\mu < 5{,}000$
2. Choose significance level: 5%
3. Look up critical value: $z^* = 1.64$
4. Calculate the test statistic: $z = -1.79$
5. Decision: reject H_0 since $-1.79 < -1.64$ and falls into the rejection region

One vs two tailed tests

- Should you use a one tailed (H_1: $\mu < 5{,}000$) or two tailed (H_1: $\mu \neq 5{,}000$) test?

- If you are only concerned about falling one side of the hypothesised value (as here: we would not worry if LEDs lasted *longer* than 5,000 hours) use the one tailed test. You would not want to reject H_0 if the sample mean were anywhere above 5,000.

- If for another reason, you *know* one side is impossible (e.g. demand curves cannot slope upwards), use a one tailed test.

One vs two tailed tests (continued)

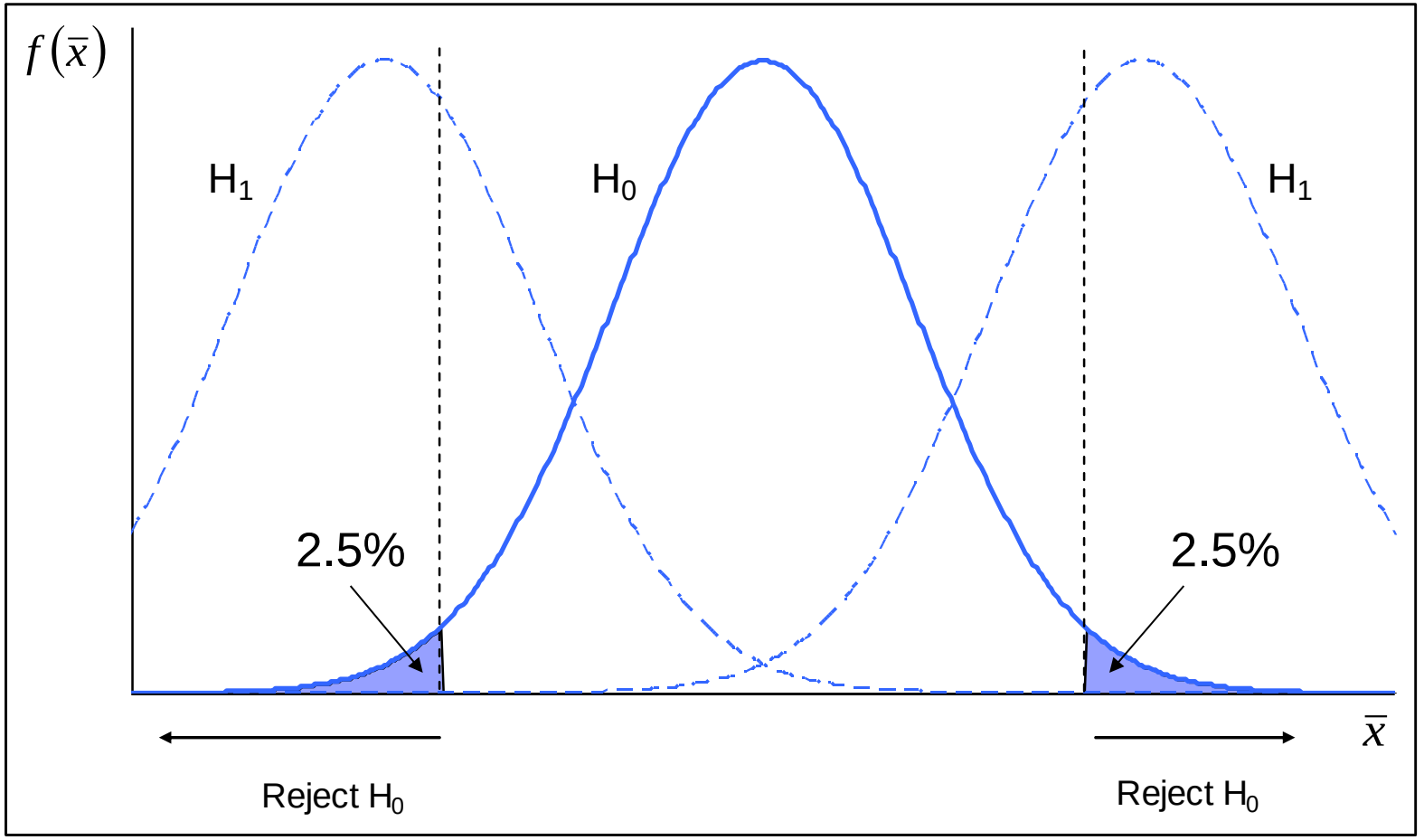
- Otherwise, use a two tailed test.
- If unsure, choose a two tailed test.
- Never choose between a one or two tailed test on the basis of the sample evidence (i.e. do not choose a one tailed test because you notice that 4,900 < 5,000).
- The hypothesis should be chosen before looking at the evidence!

Two tailed test example

- It is claimed that an average child spends 15 hours per week watching television. A survey of 100 children finds an average of 14.5 hours per week, with standard deviation 8 hours. Is the claim justified?

- The claim would be wrong if children spend either more *or less* than 15 hours watching TV. The rejection region is split across the two tails of the distribution. This is a two tailed test.

A two tailed test – diagram



Solution to the problem

1. H_0: $\mu = 15$
   H_1: $\mu \neq 15$
2. Choose significance level: 5%
3. Look up critical value: $z^* = 1.96$
4. Calculate the test statistic:

$$z = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} = \frac{14.5 - 15}{\sqrt{8^2/100}} = 0.625$$

5. Decision: we do not reject H_0 since 0.625 < 1.96 and does not fall into the rejection region

The choice of significance level

- Why 5%?
- Like its complement, the 95% confidence level, it is a convention. A different value can be chosen, but it does set a benchmark.
- If the cost of making a Type I error is especially high, then set a *lower* significance level, e.g. 1%. The significance level is the probability of making a Type I error.

The prob-value approach

- An alternative way of making the decision
- Returning to the LED problem, the test statistic $z = -1.79$ cuts off 3.67% in the lower tail of the distribution. 3.67% is the prob-value for this example
- Since 3.67% < 5% the test statistic *must* fall into the rejection region for the test

Two ways to rejection...

Reject H_0 if *either*

- $z < -z^*$ (-1.79 < -1.64)

or

- the prob-value < the significance level (3.67% < 5%)

Testing a proportion

- Same principles: reject H_0 if the test statistic falls into the rejection region.

- To test H_0: $\pi = 0.5$ vs H_1: $\pi \neq 0.5$ (e.g. a coin is fair or not) the test statistic is

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{p - 0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}}$$

Testing a proportion (continued)

- If the sample evidence were 60 heads from 100 tosses (p = 0.6) we would have

$$z = \frac{0.6 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 2$$

- so we would (just) reject H_0 since 2 > 1.96.

Testing the difference of two means

- To test whether two samples are drawn from populations with the same mean
- H_0: $\mu_1 = \mu_2$ or H_0: $\mu_1 - \mu_2 = 0$
  H_1: $\mu_1 \neq \mu_2$ or H_0: $\mu_1 - \mu_2 \neq 0$
- The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Testing the difference of two proportions

- To test whether two sample proportions are equal
- H_0: $\pi_1 = \pi_2$ or H_0: $\pi_1 - \pi_2 = 0$
  H_1: $\pi_1 \neq \pi_2$ or H_0: $\pi_1 - \pi_2 \neq 0$
- The test statistic is

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n_1} + \frac{\hat{\pi}(1-\hat{\pi})}{n_2}}} \qquad \hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Small samples ($n < 25$)

- Two consequences:
  - the *t* distribution is used instead of the standard normal for tests of the mean

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

  - tests of proportions cannot be done by the standard methods used in the book

Testing a mean

- A sample of 12 cars of a particular make average 35 mpg, with standard deviation 15. Test the manufacturer's claim of 40 mpg as the true average.

- H_0: $\mu = 40$
  H_1: $\mu < 40$

Testing a mean (continued)

- The test statistic is

$$t = \frac{35 - 40}{\sqrt{15^2/12}} = 1.15$$

- The critical value of the *t* distribution (df = 11, 5% significance level, one tail) is *t** = 1.796

- Hence we cannot reject the manufacturer's claim

Testing the difference of two means

- The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_1) - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}$$

- where S^2 is the pooled variance

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Summary

- The principles are the same for all tests: calculate the test statistic and see if it falls into the rejection region
- The formula for the test statistic depends upon the problem (mean, proportion, etc)
- The rejection region varies, depending upon whether it is a one or two tailed test