Tues 2pm.

University of BRISTOL

Applied Statistics

Lecture 4

David Barton & Sabine Hauert

Department of Engineering Mathematics

Outline

- Hypothesis testing
- Functions of random variables
- Sample means

OpenIntro Statistics

Chapter 4, particularly §4.1, and §4.3

Outline of hypothesis testing

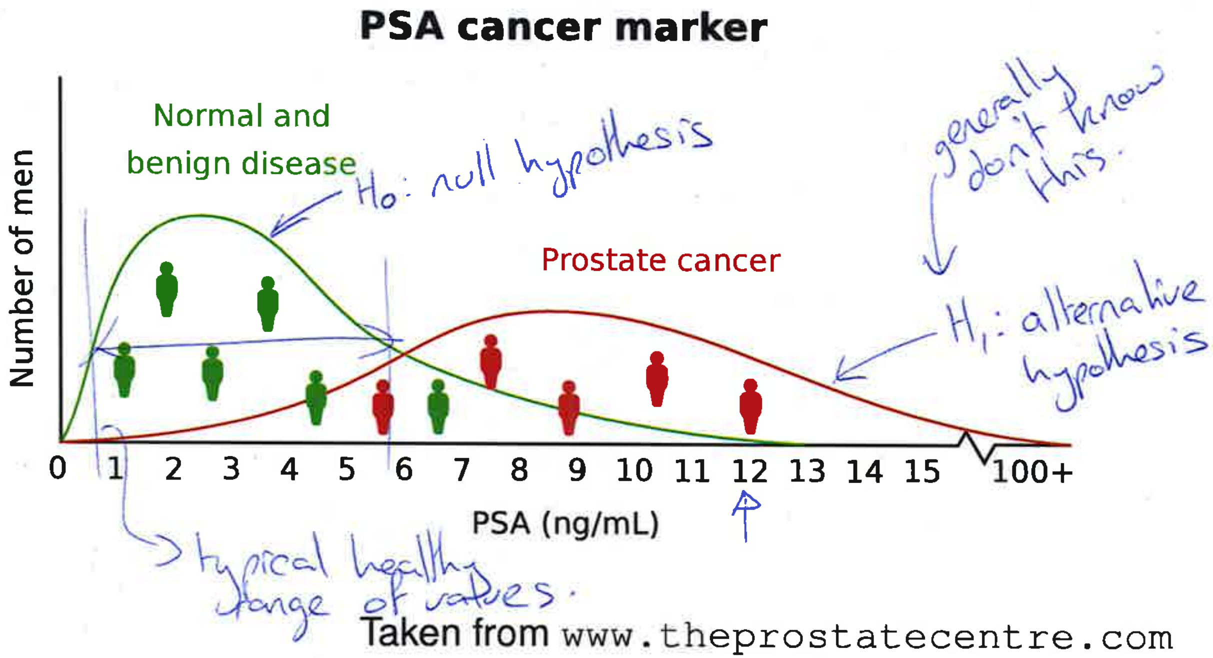
Hypothesis testing

1. Construct your null hypothesis H_0. → what you think is the [in the absence of any data.]
2. Decide on an appropriate mathematical model for H_0. → can only use data to show that H_0 is unlikely.
   - Consider the statistical assumptions being made about the sample!
3. Determine the corresponding probability distribution for H_0.
4. Decide whether it is a one-tailed problem or two-tailed.
5. Decide on the significance level that is appropriate. → 5% arbitrary significance level
6. *Look at the data* and calculate whether it is a statistically significant result. → calculate p-value

Do not look at the data before deciding on all the details of the test!

[OK, that can be a little difficult in the exercises here]

Prostate cancer testing



Taken from www.theprostatecentre.com

Beyond a sample size of one

Example

Hypothesis H_0: the daily energy output of a wind turbine is normally distributed with mean 720 MJ and standard deviation 200 MJ.

A particular turbine is measured for 10 days to give an average output of 600 MJ/day — is the model wrong?

$X \sim N(720, 200^2)$ (variance = standard deviation^2)

$N(\mu, \sigma^2)$

μ → mean, σ^2 → variance.

Summing random numbers

Take a step back: work out the distribution of the sum $Y = X_1 + X_2$ of two normally distributed variables.

Fact (Sum of two normal variables)

Given two normally distributed variables

$$X_1 \sim N\left(\mu_1, \sigma_1^2\right) \quad \textit{and} \quad X_2 \sim N\left(\mu_2, \sigma_2^2\right)$$

their sum $X_1 + X_2$ *is also normally distributed*

$$X_1 + X_2 \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

as is their difference

$$X_1 - X_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

$$\bar{X} = \frac{1}{10}\left(X_1 + X_2 + X_3 + \dots + X_{10}\right)$$

$$X_i = N(720, 200^2) \qquad \sum_{i=1}^{10} X_i \sim N\left(10 \times 720, 10 \times 200^2\right)$$

Scaling normal random numbers

The second ingredient is how to scale normally distributed random variables.

Fact (Scaling a normally distributed variable)

Given a normally distributed variable

$$X \sim N(\mu, \sigma^2)$$

multiplying it by a constant results in

$$aX \sim N(a\mu, a^2\sigma^2).$$

$$\frac{1}{10} \sum_{i=1}^{10} X_i \sim N\left(720, \frac{200^2}{10}\right)$$

$$\sum \quad X_1 + X_2 + X_2 \neq 3X_1$$

i.i.d.

The sample mean

To answer this question we need to know what the probability distribution of the mean of a sample of 10 individuals.

$X \sim N(720, 200^2)$ → population mean.

↳ μ ↳ σ^2

Definition (Sample mean)

Given a sample of n random individuals X_i drawn from a particular distribution the sample mean of those samples is

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

μ & σ^2 from H_0 constants!

Since X_i are random samples, $\bar{x}$ is a random variable.

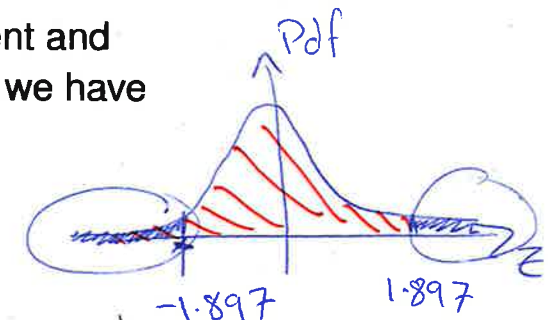
Since the sample mean $\bar{x}$ is a random variable, *it has its own probability distribution*. The population mean μ, however, is *not* a random variable.

Distribution of the (normal) sample mean

Standard Normal
$z \sim N(0,1)$

Since for the sample mean the individuals are i.i.d. (independent and identically distributed) and *in this example* normally distributed we have

$\hookrightarrow X_i \sim N(\mu, \sigma^2)$



$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

In the example, the hypothesised turbine model has $\mu = 720$ and $\sigma = 200$ and so $\bar{x} \sim N(720, 200^2/10)$ we can work out

What is $P(\bar{x} \leq 600)$? $= P\left(z \leq \frac{(600-720)}{200/\sqrt{10}}\right)$

$= P(z \leq -1.897)$

$= P(z \geq 1.897) = 1 - \underbrace{P(z < 1.897)}_{} \approx 1 - P(z < 1.90)$ $\boxed{= 0.028} \times 2$

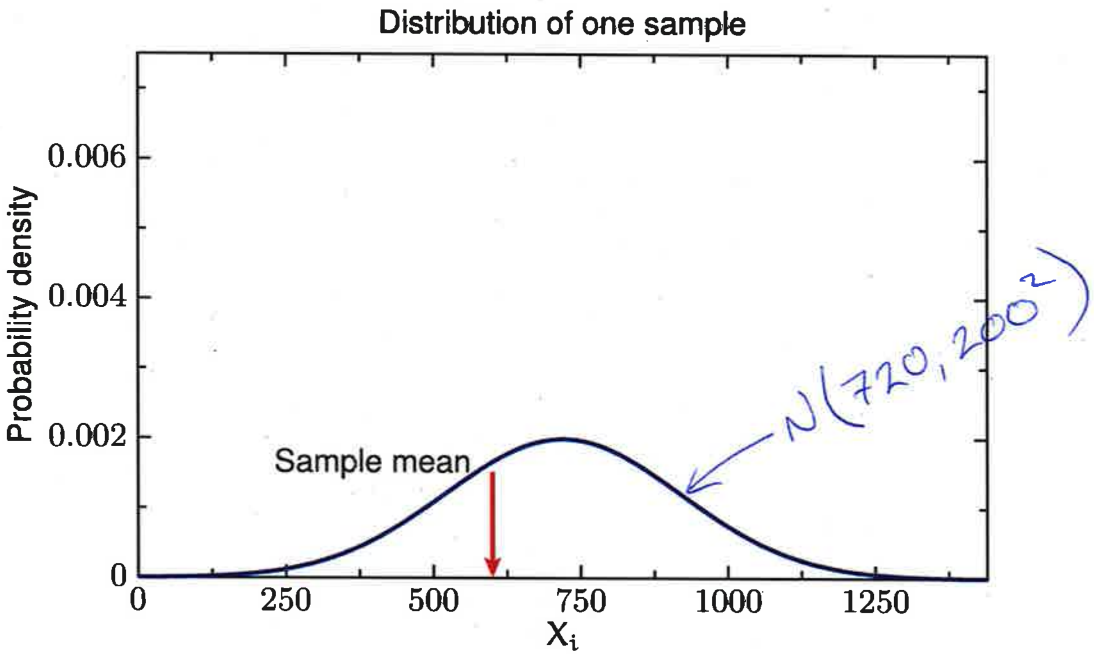
Hence $\boxed{p = 0.056}$ probability of seeing a difference of 120 between the sample mean and true mean (not significant to 5%, two-tailed test)

multiply by two to account for differences in both directions.

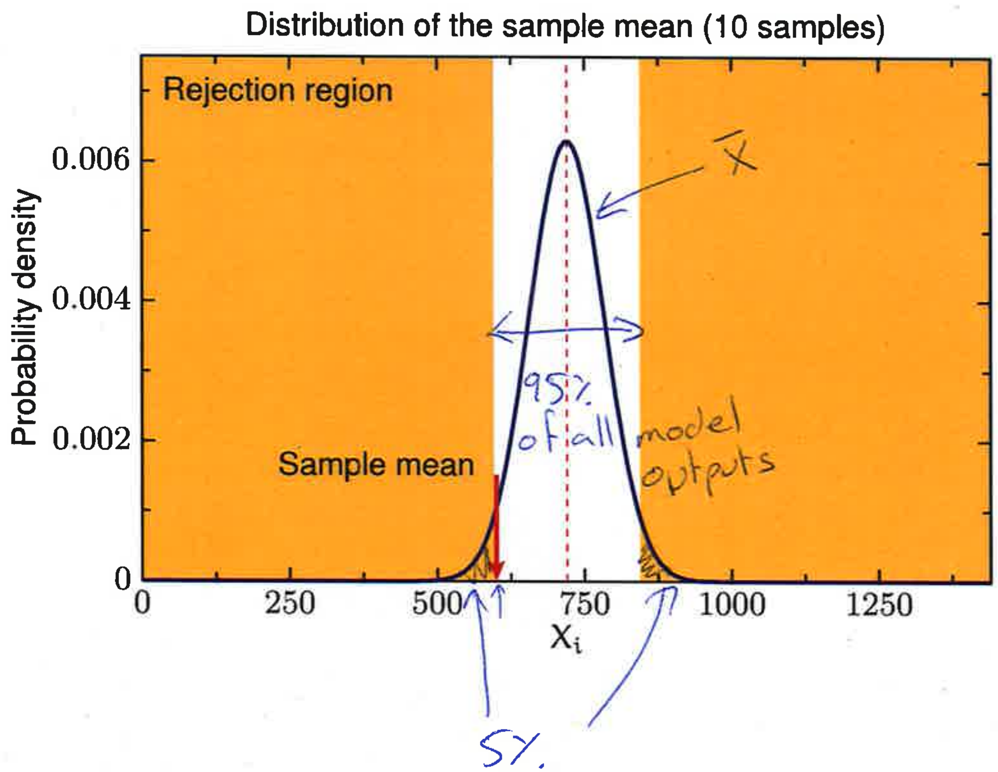
Don't reject Ho since $p > 0.05$

two-tailed test ←

Wind turbine in graphs — one individual



Wind turbine in graphs — sample mean



Take care with the problem statement

Notice how the question was reframed was changed, from

> *is the model wrong?*

to

> *is the probability of a difference of 120 between the sample mean and the hypothesised mean small?*

This is a key point in statistics — how do you go from a high-level question to a precise mathematical question that you can answer?

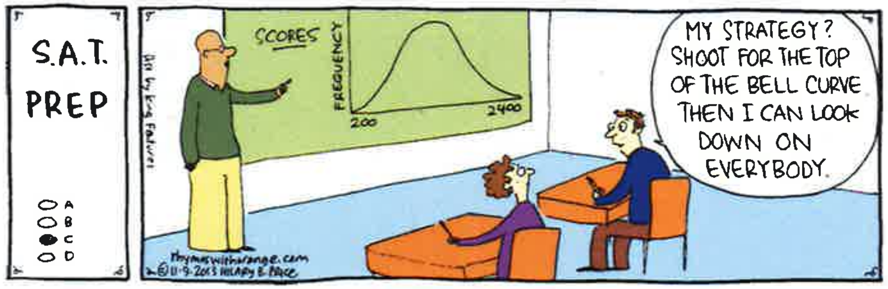
> *There are three types of lies — lies, damn lies, and statistics. (Benjamin Disraeli; disputed)*

Exercises

Quote of the day

George Bernard Shaw

Statistics show that of those who contract the habit of eating, very few survive.



Exercises

- 4.17–4.20 from OpenIntro Statistics (again)