COGS2020

TUTORIAL 6: CLT AND STATISTICAL TESTING BASICS

Quick recap

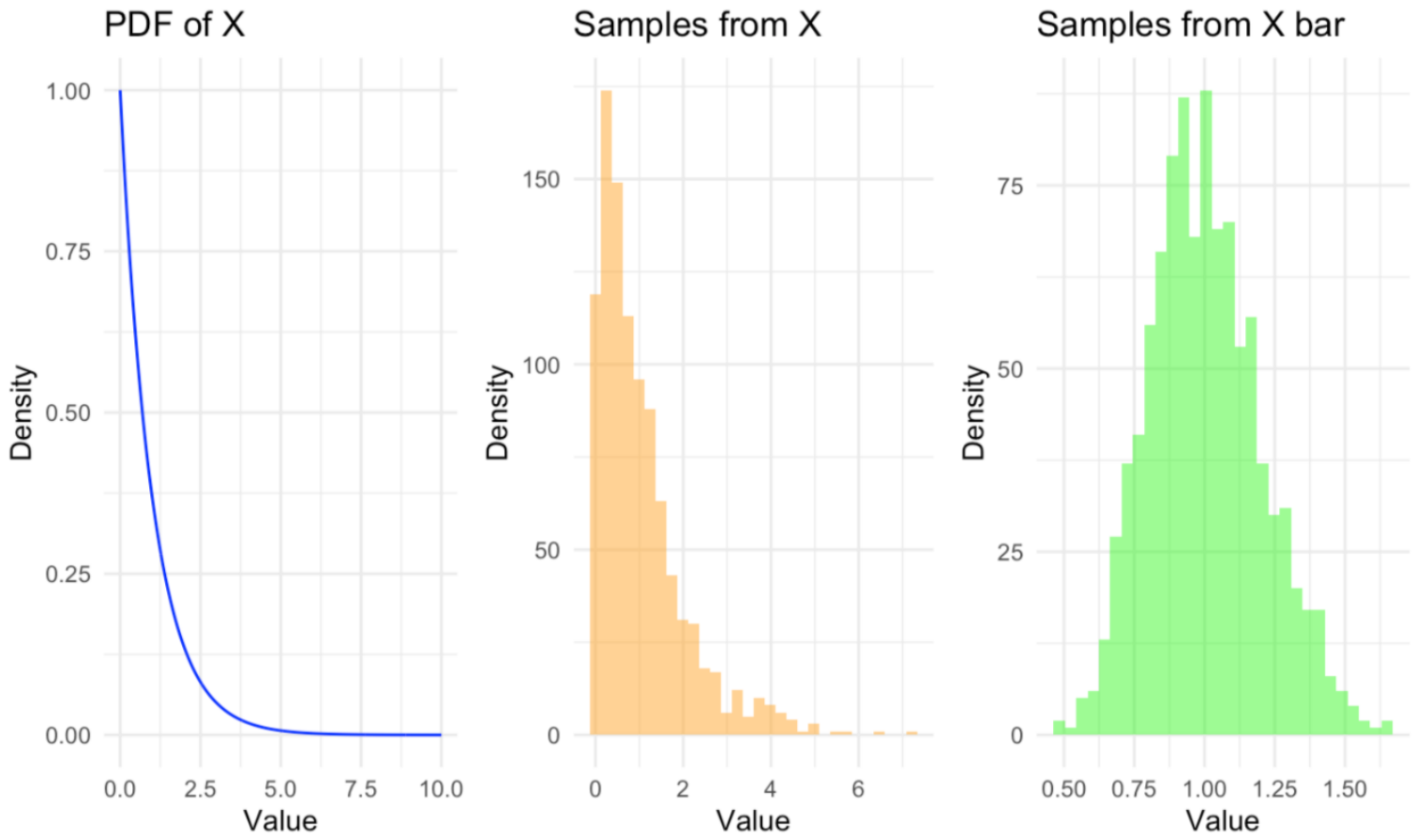
- Last week, we learned about probability basics, random variables, their probability distributions, and different functions (e.g. PDF, CDF, etc.) we can use to describe them
- We also learned that probability distributions are defined/created by identifying key moments (e.g. expected value, variance)
- Sample statistics and graphs are all estimates of the true population

Central Limit Theorem

- Definition: As **n increases** (rule of thumb, at least **n = 30**), the **distribution of sample means** (or any other linear transformation) will **approximate a normal distribution** regardless of the original population distribution the samples were drawn from
  - Assuming that the sample means are averages of **independent and identically distributed (i.i.d) random variables**

- Why does the CLT matter?
  - This theorem underlies all null hypothesis testing – will come back to this in a few slides, but first, let's look at an example and break down Matt's math in his lecture slides

Matt's example



Key things to notice:

1. PDF shows the probability density of original population distribution – **high probability at lower numbers**
2. Shows a histogram of an **actual sample** – shape follows the original distribution as expected
3. **Sampling distribution** – if we were to take multiple samples, then get the mean of each sample, and make a distribution.. **Distribution of sample means**, is a ***normal distribution***

This shows us exactly what the CLT states – the distribution of sample means will appropriate a normal distribution regardless of the original population distribution

Matt’s maths

- In short – we are illustrating the CLT in mathy terms. We are going to look at how we can transform random variables, and how that effects key moments such expected value and variance.
- Let’s look at the CLT applied to different transformations – in means, sums, etc.

Distribution of Sample Means

The RV that generates means of X

$$\boldsymbol{Y} = \frac{1}{n}(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n)$$

i.i.d RVs – the RVs that generate data from our population

This equation just shows us how X is transformed into Y – it shows us how doing things to our random variables X can produce another random variable we arbitrarily call Y

- The key moments of Y can be expressed in terms of X – remember Y is defined by X (+ some transformation)

$$\begin{aligned} E(\boldsymbol{Y}) &= E(\frac{1}{n}(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n)) \\ &= \frac{1}{n}(E(\boldsymbol{X}_1) + E(\boldsymbol{X}_2) + \cdots + E(\boldsymbol{X}_n)) \\ &= \frac{1}{n}(\mu_x + \mu_x + \ldots + \mu_x) \\ &= \mu_x \end{aligned}$$

$$\begin{aligned} Var(\boldsymbol{Y}) &= Var(\frac{1}{n}(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n)) \\ &= Var(\frac{1}{n^2}(Var(\boldsymbol{X}_1) + Var(\boldsymbol{X}_2) + \cdots + Var(\boldsymbol{X}_n)) \\ &= \frac{1}{n^2}(\sigma_x^2 + \sigma_x^2 + \cdots + \sigma_x^2) \\ &= \frac{1}{n}\sigma_x^2 \end{aligned}$$

Distribution of Sample Sums

The RV that generates sums of X

$$\boldsymbol{Y} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n$$

i.i.d RVs – the RVs that generate data from our population

This equation just shows us how X is transformed into Y – it shows us how doing things to our random variables X can produce another random variable we arbitrarily call Y

- The key moments of Y can be expressed in terms of X – remember Y is defined by X (+ some transformation)

$$\begin{aligned} E(\boldsymbol{Y}) &= E(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n) \\ &= E(\boldsymbol{X}_1) + E(\boldsymbol{X}_2) + \cdots + E(\boldsymbol{X}_n) \\ &= \mu_x + \mu_x + \ldots + \mu_x \\ &= n\mu_x \end{aligned}$$

$$\begin{aligned} Var(\boldsymbol{Y}) &= Var(\boldsymbol{X}_1 + \boldsymbol{X}_2 + \cdots + \boldsymbol{X}_n) \\ &= Var(\boldsymbol{X}_1) + Var(\boldsymbol{X}_2) + \cdots + Var(\boldsymbol{X}_n) \\ &= \sigma_x^2 + \sigma_x^2 + \cdots + \sigma_x^2 \\ &= n\sigma_x^2 \end{aligned}$$

General Rule

$$\text{Let} Y \sim aX + b$$
$$\mathbb{E}[Y] = a\mathbb{E}[X] + \mathbb{E}(b)$$
$$\mathbb{V}\text{ar}[Y] = (a^2)\mathbb{V}\text{ar}[X]$$

- These equations are like a general rule used to describe linear transformations of a random variable X (into Y), and how that effects key moments like expected value and variance (of Y)
- This rule underlies the E[X] and Var[X] equations we saw for the sample sums and average random variables
- But don't worry – I will not go into the math about how exactly we are able to get there from these rules ☺

Interim summary + why does this all matter?

- We have essentially shown that when we create a distribution of sample means (Y), we can define its key moments/descriptives using the original data from the population (X)

$$\mu_Y = \mu_X$$

$$\sigma_Y^2 = \frac{1}{n}\sigma_X^2$$

Note: Y can also be notated as $\bar{X}$ (big X bar)

- In summary, the CLT gives us 3 main implications:
  1. **Sampling distribution of sample means will ALWAYS be normal** (as long as n > 30) regardless of the original distributions
  2. The **mean of our sampling distribution will be equal to the true population mean**
  3. The **variance of our sampling distribution will always be LESS than our true population distribution**
- Importance of the CLT: we *LOVE* normal (ish) distributions in research, we know a lot about them, meaning that we can use the normal (ish) distribution to model things (e.g. the null hypothesis)

Using sampling distributions to represent our population parameters

- This is exactly what we have been doing so far – we will just be changing the notation a little bit

$$Y = \overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Y is the same thing as $\bar{x}$, the RV that generates sample means

X the RV that represents our population – this gives us our raw data

Sampling distribution parameters

$$\mu_{\overline{X}} = \mu_X$$

$$\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$$

Population parameters

Moments of our sampling distribution ($\bar{x}$) can be used to describe/estimate moments of the original population distribution (X)

Experimentation and estimating population parameters

- When we run experiments, we are trying to get a mini snapshot of the wider phenomenon/population – we collect a sample
- The statistics/graphs we get from these experiments (e.g. sample mean, sample variance, histogram plots, etc.) are all estimates of the wider true population, so:

The population statistics we are estimating

$$\hat{\mu}_X = \bar{x}$$

$$\frac{\hat{\sigma}_X}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

Sample statistics

Interim summary 2 + next step

- We have set up the stage for null hypothesis testing:
  - We know that sample statistics are all estimates of the true population
  - We have a way to model our population using a sampling distribution (and we know this is viable because of the CLT)

- “Modelling” the population is the basis of null hypothesis testing.. let’s shift gears into what null hypothesis testing is first

Null hypothesis testing

- It is a statistical method used to test an assumption about a population parameter (we assume the null)
- In null hypothesis testing, we are trying to see if there is evidence that suggests there is an effect or not
  - Example 1: We want to see if the average height of MQ students is significantly greater than 160cm?
  - Example 2: Is the firing rate of a neuron is significantly greater than baseline?
- How do we do this?
  - By creating a hypothetical null model, then comparing our observed statistics (e.g. sample mean) to the null..

How do we create a hypothetical null model?

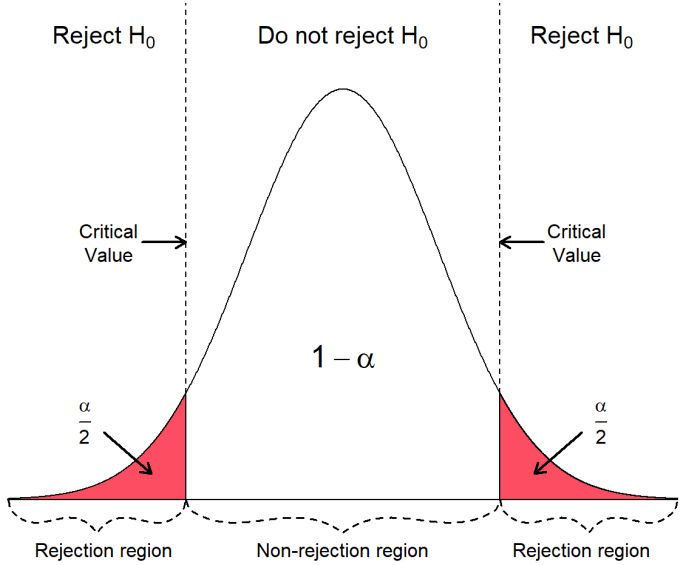
1. **Define the null and alterative** (what is considered no effect in the population?). (H0 and H1 in terms of the population parameter θ)
2. **What does the null look like in the population** (what would the population look like **if null is true** – define the key moments)

   $X \sim N$(parameters if null was true)
3. **How are we estimating the population mean (θ-hat)?** (eg. sample mean)
4. What is the **sampling distribution of this estimate** (given null is true)? **Building the sampling distribution null model**..

   $\bar{X} \sim$ *distribution*(sampling dist parameters is null was true)
5. Do we have enough information about the population to create this null model distribution? If not, what alternative distribution could we use?
6. What is the **sample/test statistic ($\theta\text{-hat}_{obs}$)** based on the distribution you use to model the null?
7. **Where does sample/test statistic lie when put into your null model**? Is it likely to occur or unlikely?
   - If likely – do not reject null
   - If unlikely – reject null

Likely vs unlikely to occur – what does this mean?

- The distribution you used to model your null hypothesis is a probability distribution.
- E.g. in a normal test, you use the normal distribution to model what the null would look like in a distribution of sample means
- We know a lot about normal distributions



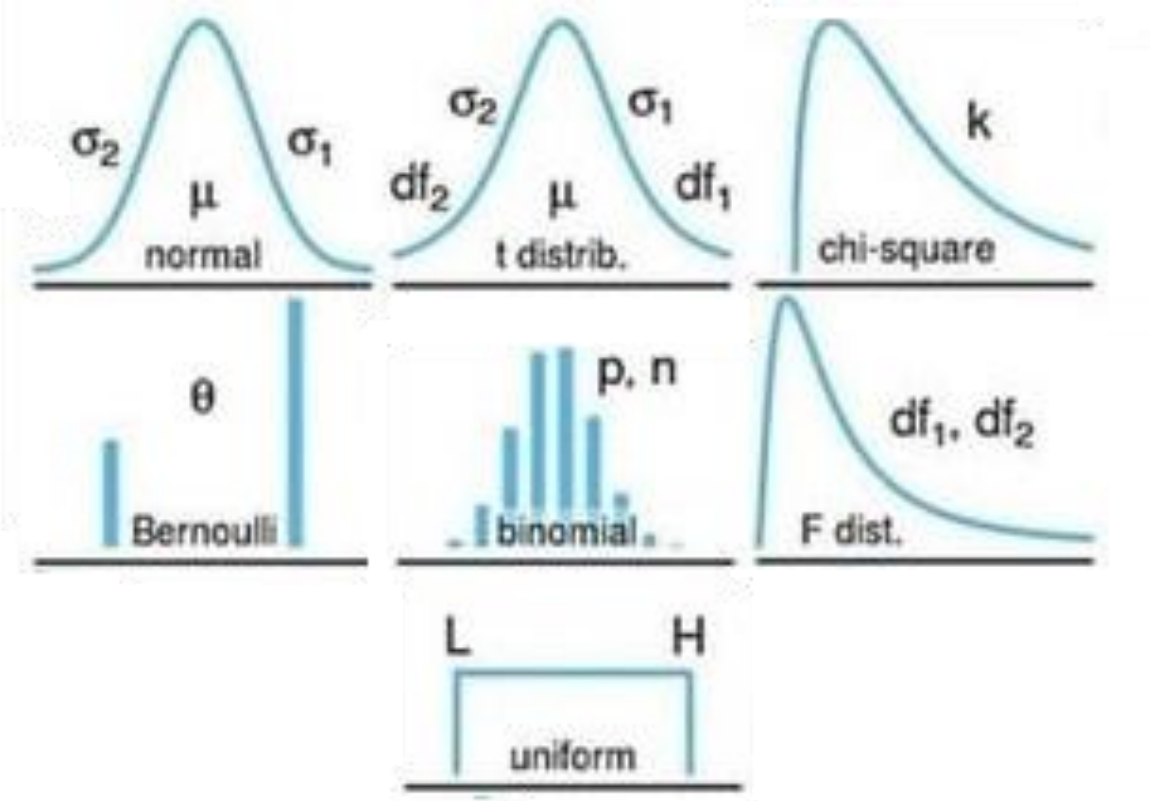
By convention, we consider the extreme 5% to be unlikely. That is why we set our alpha to be 5% or 0.05.

This is our **"rejection" zone**. If our observed sample mean falls within this area, we consider it unlikely that we got this value in our null model. Therefore, we reject the null model. We assume that it is *more likely* for another model to have produced our sample mean.

SLIDE FROM WEEK 5

Some Types of Probability Distributions

- Probability distributions are defined by specific parameters that dictate the distribution's shape/form

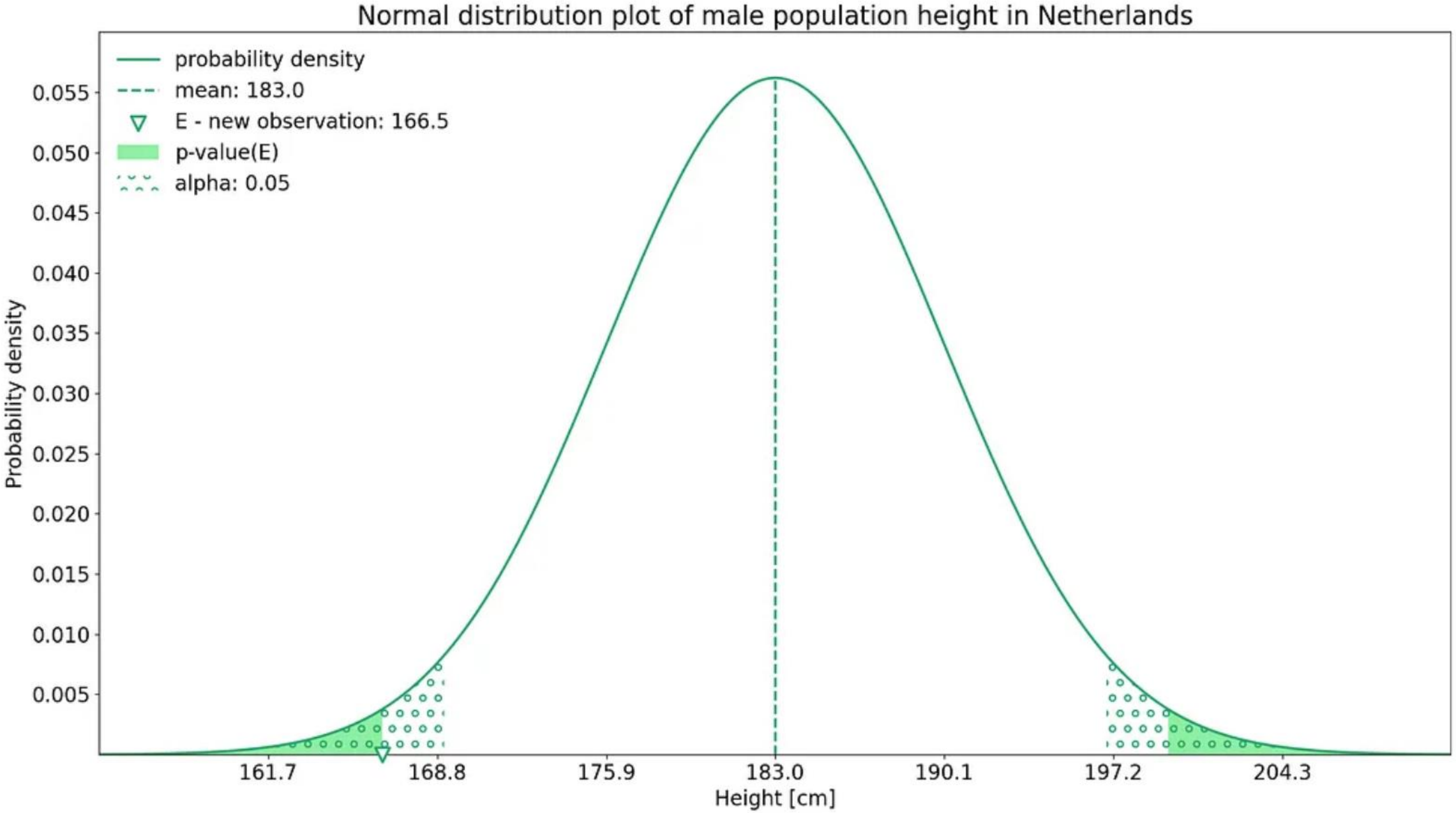


Note: x axis is sample space
y axis is probability

Random variables and their defining probability distributions are usually notated like this:

$X \sim Binomial(n,p)$
$X \sim N(\mu_X, \sigma_X^2)$
$X \sim t(df)$
$X \sim F(d_1, d_2)$

We set up our rejection zones in these distributions too – this unit only covers some of these


Normal distribution plot of male population height in Netherlands
probability density
mean: 183.0
E - new observation: 166.5
p-value(E)
alpha: 0.05
Probability density
0.055
0.050
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
161.7
168.8
175.9
183.0
190.1
197.2
204.3
Height [cm]

Alpha, p values, critical values, confidence intervals – different sides of the same coin

- **P value** = area under the curve corresponding to the probability of getting your sample/test statistic or more extreme
- **Alpha** = area under the curve representing your rejection zone (usually extreme 5%), values considered unlikely
- **Critical value(s)** = the cut-off value (on your null model) that marks where your rejection zone starts
- **Confidence interval** = the inverse of alpha, the range of values that are considered likely in your null model
  - These are sample specific values – no longer probability-based, moves back into the sample data realm

So…… what now?

- We have zoomed through all the basics – these slides are a (very short) summary and alternative explanation of Matt's lectures
- Any questions/clarifications so far?
- If you have specific questions about lecture content (e.g. one-tail vs two-tail, etc.) please ask away and/or come up to me to have a chat
- I am also happy to go through more examples – will have to be my whiteboard drawings
- Otherwise – use this time to do Problem Set 2 or get started on the final project