SO5041 Unit 8: The Normal Distribution

Brendan Halpin, Sociology

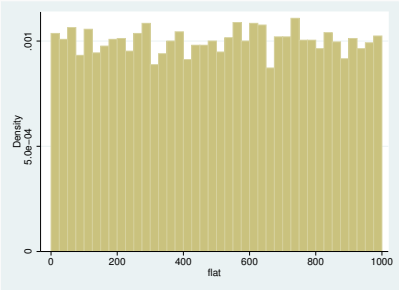
Autumn 2023/4

SO5041 Unit 8

Characteristics of distributions

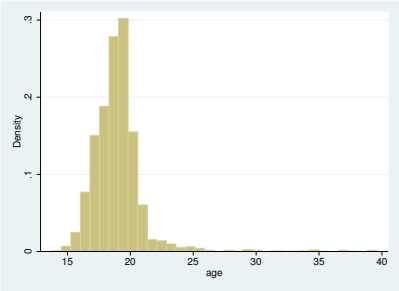
Histograms and distributions

- Histograms display distributions
- Probability distributions describe them formally
- The set of ticket numbers in a raffle has a uniform distribution
  - flat histogram
  - equal numbers of tickets in all ranges, e.g., 1–100, 101–200, 201–300
  - Thus winning ticket is equally likely to fall in any range: number is not related to chance of selection



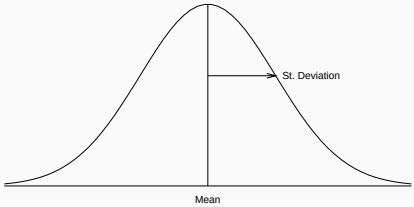
Heaped histograms

- However, if we were to pick a school-leaver at random from the school-leaver's survey, there is a relationship with date of birth:
  - ages near 19 more likely
  - ages much younger or much older much less likely
- . . . a clustered distribution



The Normal Distribution

- The extent of clustering depends on shape, standard deviation
- Smaller standard deviation means individual chosen at random more likely to fall near the mean
- The Normal Distribution a kind of “ideal type” of clustered symmetric distribution (arises naturally in many contexts)

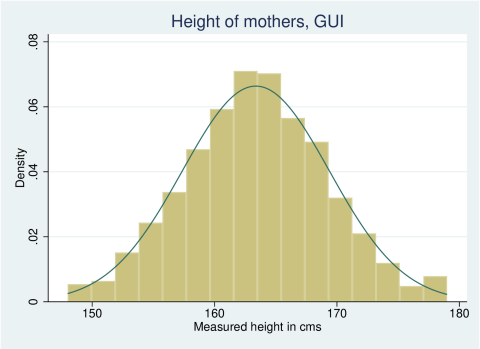


Normal

- The normal distribution is
  - symmetric
  - uni-modal
  - meso-kurtic
  - range $-\infty$ to $+\infty$
  - continuous, not discrete
- Completely defined by its mean and standard deviation:
  http://teaching.sociology.ul.ie:3838/apps/normsd/

Examples of normally distributed variables

- IQ scores and other standardised tests – designed that way
- Time for a Deliveroo – possibly normal, e.g., mean of 30 mins, standard deviation of 5 mins
- Adult human height (separately for males and females)



Why does the Normal Distribution crop up so often?

- Where there is a core value, but lots of small things pushing it either way
- First observed in physical measurements, e.g., height of mountain, speed of light
  - Unknown correct answer
  - Each measurement full of small factors (errors) pushing it up and down
  - Some errors cancel each other, some compound
- Measurements will tend to have a normal distribution (hopefully) centred on the true value
- Normally distributed if many small factors, pushing up equally to down
  - additive
  - independent

Visualisations

https://commons.wikimedia.org/wiki/File:Galton_box.webm

https://teaching.sociology.ul.ie/so4046/quincunx.mp4

It crops up in sampling

- Each case in a sample pulls the sample mean up and down
  - For calculating things like means, each case has an additive effect
  - In simple random samples each case is independent of all others
- Therefore the set of all sample means has a normal distribution:
  http://teaching.sociology.ul.ie:3838/so4046/sampling/
- We will come back to this in the next lecture

SO5041 Unit 8

The Standard Normal Distribution

Standard Normal Distribution

- The Standard Normal Distribution is a special case of the normal distribution:
  - Mean = 0
  - Standard deviation = 1
- We can map any given ND onto it by
  - subtracting the mean
  - dividing by the standard deviation
- We can thus use the SND to estimate probabilities/proportions for any normal distribution, once we know the mean and standard deviation

Standard normal app

http://teaching.sociology.ul.ie:3838/apps/snd

Reading the SND in the app

- The app tells us the
  - proportion of the distribution, or equivalently,
  - the chance of picking a case at random
- above or below a certain value
- or inside or outside +/- times that value
- we can also use it to calculate proportions/probabilities between (or outside) any two values

Proportion above or below a given level

- For example, given mean 100 and standard deviation 20, what's the chance of observing a value above 130?

$$\mu = 100, \sigma = 20, X = 130$$

$$\Rightarrow z = \frac{X - \mu}{\sigma} = \frac{130 - 100}{20} = 1.5$$

- From the app, we see that z=1.5 corresponds to p=0.0668 or about 6.7%: 6.7% of the distribution is above 130
- Clearly, 100% - 6.7% = 93.3% of the distribution is below 130

Height of GUI mothers

```
. su hwomen
```

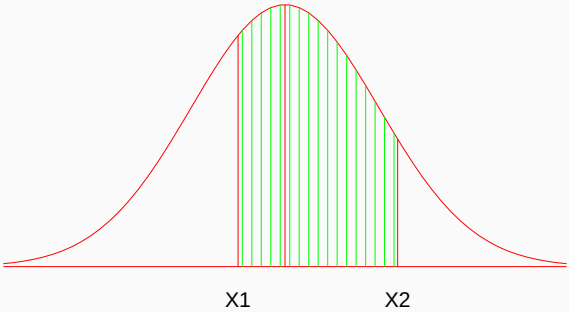
Variable	Obs	Mean	Std. dev.	Min	Max
hwomen	6,300	163.373	6.01054	148	179

```
. count if hwomen >= 175
  229

. count if hwomen >= 170
  949

. count if hwomen <= 160
  2,018

. count if hwomen <= 155
  626

. centile hwomen, centile(75 90)
```

Variable	Obs	Percentile	Centile	Binom. interp. [95% conf.	interval]
hwomen	6,300	75	167	167	168
		90	171	171	171

Proportion between two values

- Once we know how to calculate the proportion of the distribution above or below any value, we can calculate the proportion between any pair of values



Proportion between two values

- To calculate the proportion between X1 and X2, calculate
  - The proportion below X1: $P(X < X1)$
  - The proportion above X2: $P(X > X2)$
  - $P(X1 < X < X2) = 1 - P(X < X1) - P(X > X2)$

Working backwards: given p find z

- We may also wish to work in the opposite direction
- Instead of asking what proportion of the distribution is above X, we may ask what is the level such that proportion p of the distribution is above it?
- For example, given the same distribution, what is the level such that only 5% of the distribution is above it?

Working backwards: given p find z

- Given the same distribution, what is the level such that only 5% of the distribution is above it?
- We work backwards, starting in the body of the table by searching for the value nearest to 5% or 0.050
- This corresponds with z = 1.645 (falls between 1.64 and 1.65)
- Reverse the formula: $X = \sigma \times z + \mu = 20 \times 1.645 + 100 = 132.9$
- Therefore, 5% of this distribution is above 132.9

SO5041 Unit 8

Online apps

App: P<X

Find the proportion below X, (above or below the mean)

- Link: Proportion below X

App: P>X

Find the proportion above X, (above or below the mean)

- Link: Proportion above X

App: P between X1 and X2

X1 and X2 may both be on either side of, or straddle the mean.

- Link: Proportion between X1 and X2