SO5041 Unit 8: More on Distributions

Brendan Halpin, Sociology

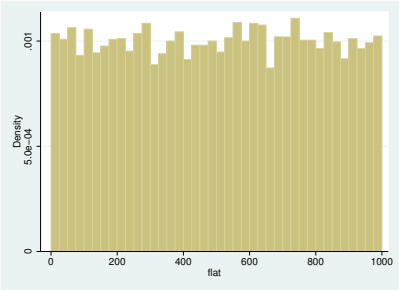
Autumn 2020/1

SO5041 Unit 8

Characteristics of distributions

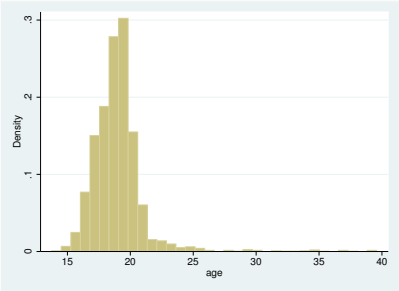
Histograms and distributions

- 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



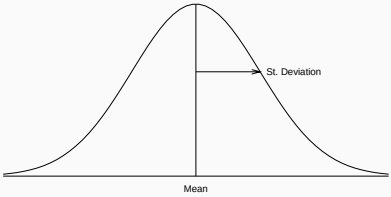
Heaped histograms

- 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



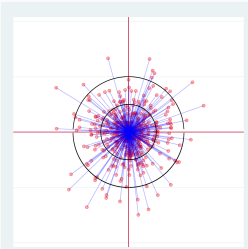
Shape and standard deviation

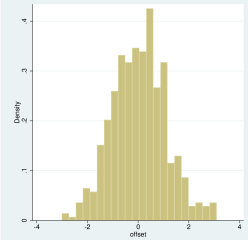
- The extent of clustering depends on shape, standard deviation
- Smaller standard deviation means individual chosen at random more likely to fall near the mean
- The Normal Distribution a kind of “ideal type” of clustered symmetric distribution



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
- Alcohol consumption of non-abstainers, e.g., mean of 20 units per week, standard deviation of 7 units
- How far darts hit relative to their target





Mean and St Deviation are enough

- Combining the mean, standard deviation and the “knowledge” that the distribution is normal means we can judge
  - the proportions above, below or between any given values
  - the chance of a case chosen at random of falling in any given range

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

Visualisations

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

https://vimeo.com/379129300

It crops up in sampling

- Each case in a sample pulls the sample mean up and down
- Therefore the set of all sample means has a normal distribution

http://teaching.sociology.ul.ie:3838/so4046/sampling/

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 Distribution Table

Table of the Standard Normal Distribution

Right tail (probability of $X > z$)

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.00	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.50	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.00	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.50	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.80	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.30	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.40	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064

Reading the table

- The table is best suited to telling us the
  - proportion of the distribution, or equivalently,
  - the chance of picking a case at random above a given value above the mean
- This is because it starts at zero (the mean of the SND) and goes up only
- However, since the normal distribution is symmetrical we can use this table for values below the mean too

Proportion above or below a given level above the mean

- 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 table, 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

Proportion below a given level below the mean

- We use the symmetry to make calculations below the mean
- For example, for the same distribution, what's the chance of observing a value below 85?

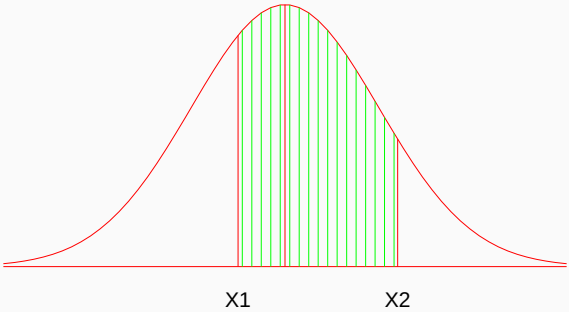
$$\mu = 100, \sigma = 20, X = 85$$

$$\Rightarrow z = \frac{X - \mu}{\sigma} = \frac{85 - 100}{20} = -0.750$$

- By symmetry, the chance of being below -0.75 is exactly the same as that of being above +0.75, so we read that:
- z = 0.75 $\Rightarrow p = 0.2266$, which means 22.7% of the distribution is below 85 (and 77.3% = 100% - 22.7% is above 85)

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