Data Exploration – Statistics 2

Dr. Markus Hofmann

Inferential Statistics

Overview

- If we are using statistics to determine something about the original population then we are talking about Inferential Statistics
- This is because we are inferring something about the population from the sample
- Note: we are going further here than describing the sample!!!
- We are attempting to determine some level of detail about objects from the population that we have NOT sampled extensively!

Overview

- When randomly selecting observations (examples) a number of times then we would obtain different samples and therefore also different measures. This is known as the **sampling error**.
- If we select many random samples then we can expect that most of the means of the random sample would in fact be close to the actual mean
- In fact, the selection of many random samples follows a normal distribution for sample sizes greater than 30

Overview

- We can further expect that increasing the sample size will result in the average mean will be closer to the actual mean. In fact the larger the sample sizes the narrower the distribution
- The relationship between the variation of the original variable and the number of examples in the sample to the sampling distribution is summarised by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

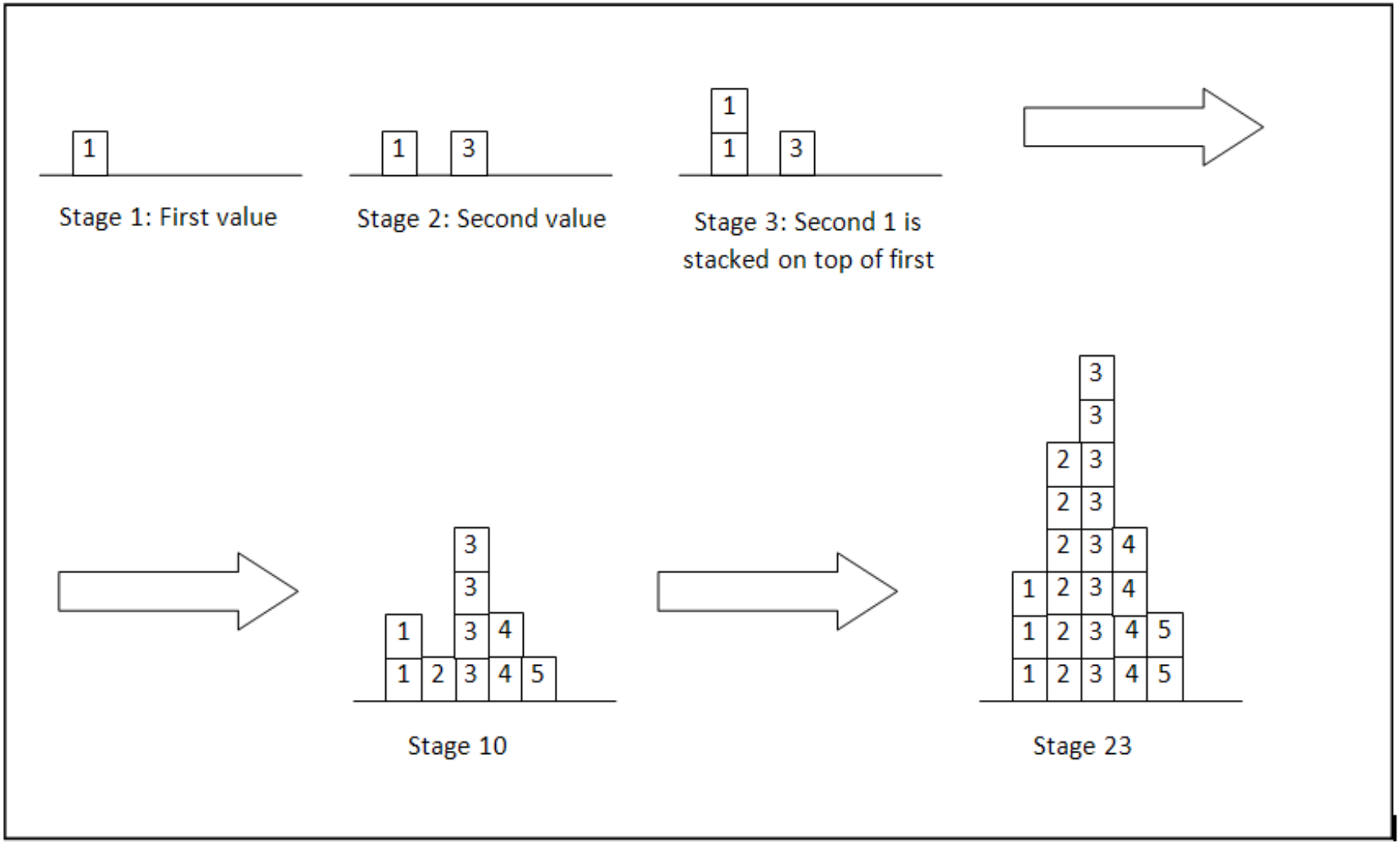
- Where $\sigma_{\bar{x}}$ is the standard deviation of the sample means, σ is the standard deviation of the population and n is the number of examples in the sample

Overview

- Therefore, when the number of sample examples increase, the standard deviation of the sample means decreases.
- The standard deviation of the sample means is also known as the Standard Error of the Mean
- The main idea behind the Standard Error statistic is that you have a sample statistic and want to build a frequency distribution of all the possible samples. Then, you want to describe the variability of these samples.
- The larger the sample, the smaller the SE.
- In inferential statistics we use exactly the sampling distribution to assess the chance or probability that we will see a range of average values.

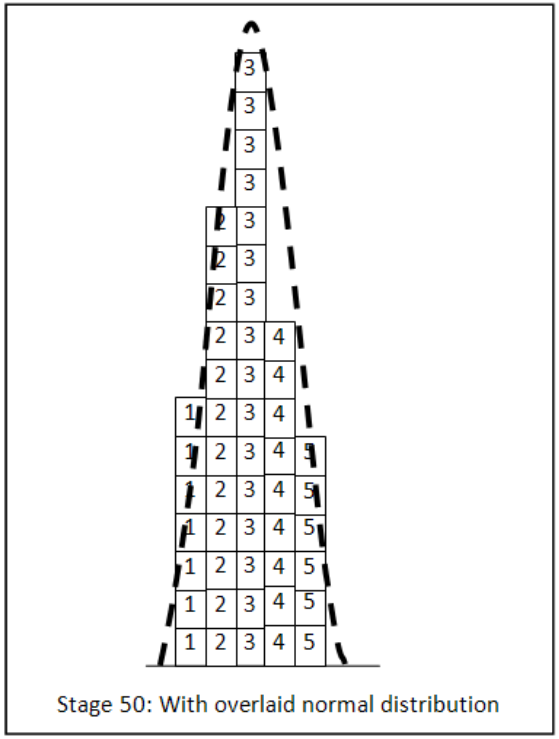
Distributions

- This is an important concept to understand
- Imagine collecting a large set of values. Then take these values (in the order they were collected) and gather then into groups, stacking values in the same group on top of each other. As the values are sorted into their respective groups you would get (for example, and in crude terms):


1
Stage 1: First value
1 3
Stage 2: Second value
1
1 3
Stage 3: Second 1 is stacked on top of first
3
3 4
1 2 3 4 5
Stage 10
3
3
2 3
2 3
2 3 4
1 2 3 4
1 2 3 4 5
1 2 3 4 5
Stage 23

Distributions

- The shape of this is called (graphical) distribution
- It tells us something about the characteristic of the data (e.g. in terms of shape, spread and centre measures)
- Very often we end up with a distribution called the "Normal Distribution"
- This might seem unrealistic but is in fact commonly a true representation of the "real world"
- The fatness of the tails is **controlled by** a parameter called the **degrees of freedom (df)**.



Stage 50: With overlaid normal distribution

Other Distributions

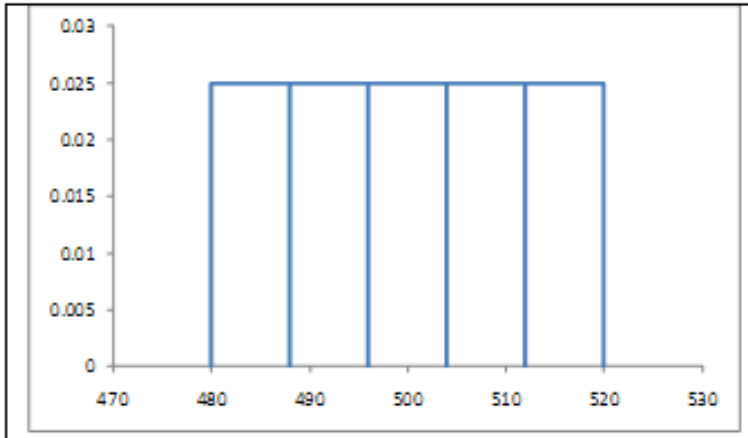


Figure 9: Uniform distribution.

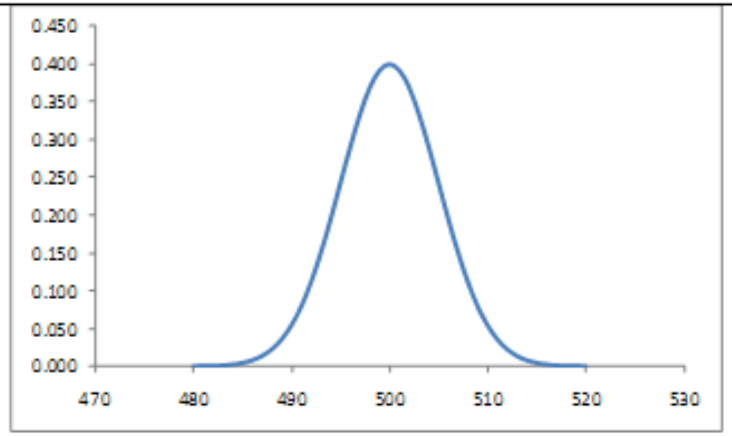


Figure 10: Normal distribution

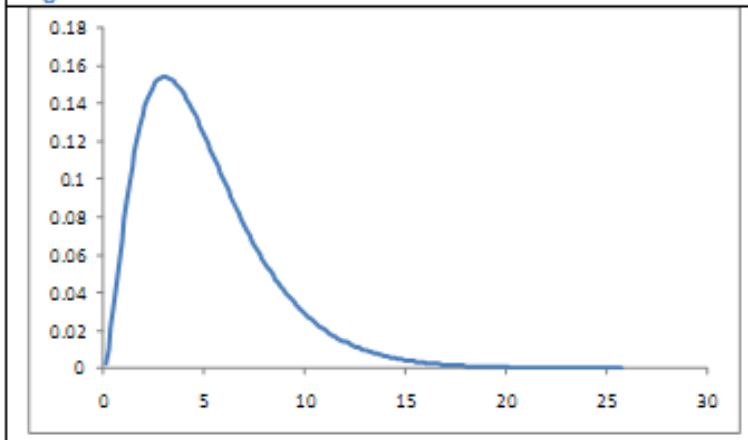


Figure 11: Chi-Squared distribution (5 degrees of freedom[7]).

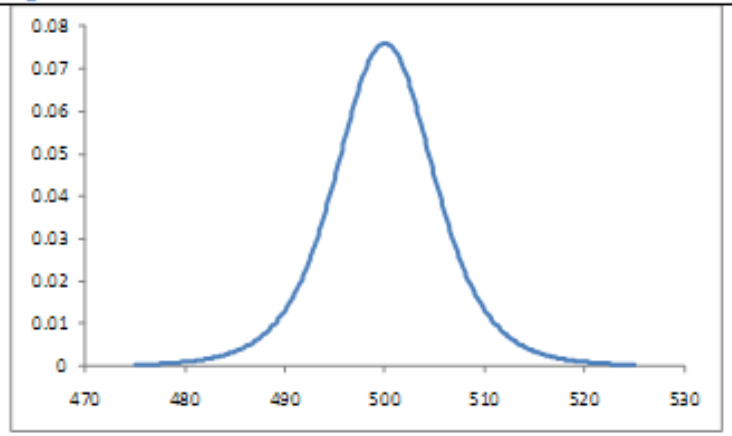


Figure 12: t-Distribution (5 degrees of freedom).

Example to Inferential Statistics

- Manufacturer wants to make a claim that the average bag of sweets contains more than 200 pieces. 500 bags are collected as sample. The average number of sweets per pack is calculated to be 201 with a standard deviation (s) of 12.
- We need to asses the probability that this value is greater than 200 or whether the difference is simply attributable to the sampling error. Let's use the sampling distribution to make this assessment.
- The area under the curve can be used to assess probabilities. 1 indicates the event will happen, 0 indicates the event will never happen. Values between those two extremes indicate the likelihood of the event to happen

Example to Inferential Statistics

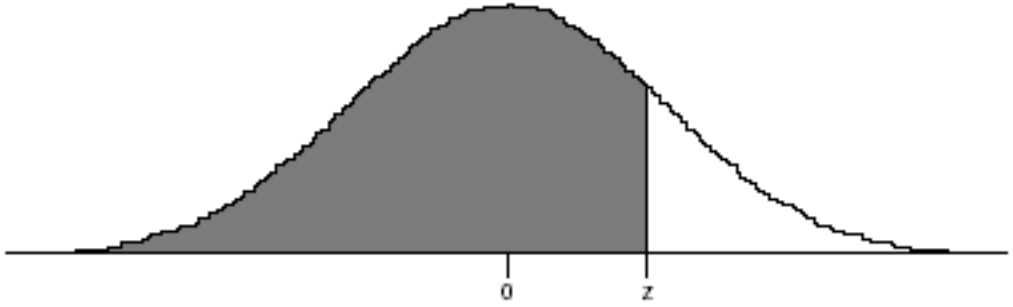
- The area under the curve of a normal distribution is 1
- The area between specific z-score ranges represents the probability that a value would lie within this range.
- We first calculate the Standard Error of the Mean using the sample standard deviation of 12 and the sample size of 500

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{500}} = 0.54$$

- To understand how many standard deviations the value 201 is away from the mean we must convert the value into a z-score

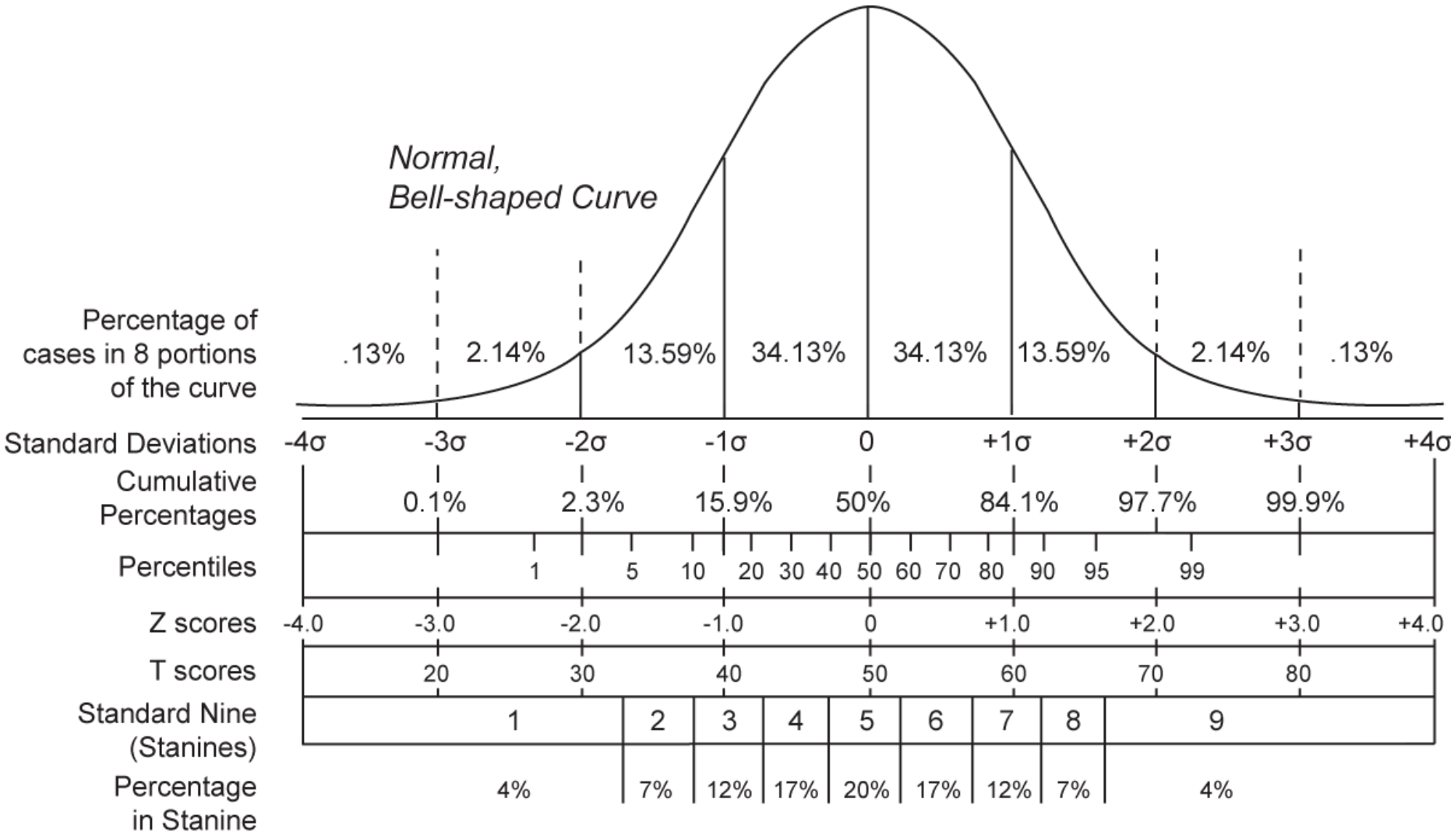
$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{201 - 200}{0.54} = 1.85$$

Normal Distribution Table



Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014

Normal Distribution and Scales



Example to Inferential Statistics

- If we plot this value we see that the value of 1.85 is larger than the mean (located to the right of the mean).
- We will later have a look how we can formalise the claim that on average there are in fact more than 200 sweets in a packet.

What's next...

- We will look at
  - Confidence Intervals
  - Hypothesis Test
  - Chi Square
  - One way analysis of variance

Confidence Interval

- A single statistic could be used as an estimate for a population (known as point estimate).
- However, we would not know the confidence of this statistic to be correct.
- Maybe we identified a reasonable confidence that the number of sweets in a packet is between 198 and 202.
- This range is known as the **Confidence Interval (CI)**.
- The CI is dependent on
  - the size of the sample
  - the required degree of confidence required

Confidence Interval

- Increasing the degree of confidence requires an increase of the range of values
- The required degree of confidence is based on the confidence level at which the estimate is to be calculated
- Let's have a look how this is done for
  - Continuous Variables
  - Categorical Variables

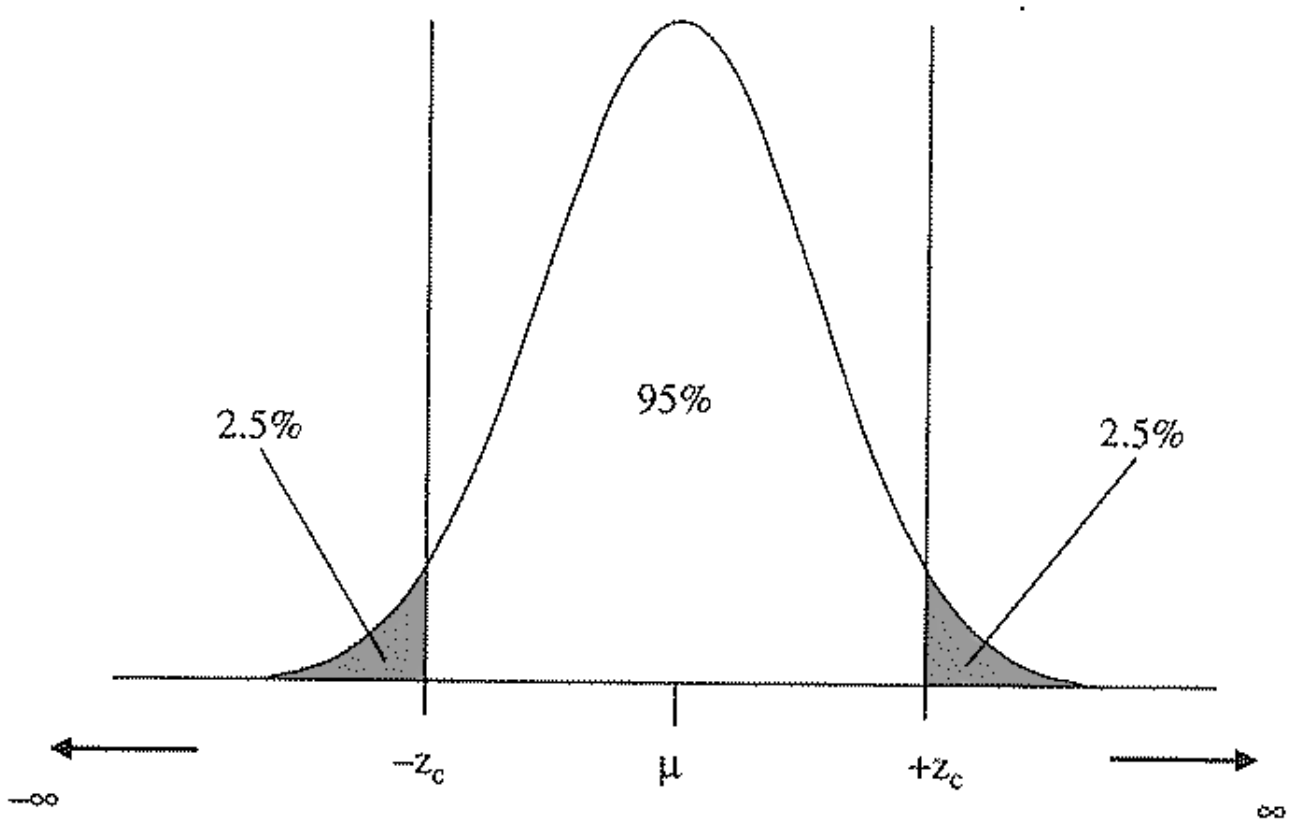
Continuous Variables

- The mean is the most common population estimate for continuous variables
- In order to determine the confidence interval we need to calculate the mean of the sample first
- The CI depends on
  - the Standard Error of the Mean
  - The required confidence
- Formula for a large sample (>30 examples): $\bar{x} \pm z_c \frac{s}{\sqrt{n}}$
- Where x bar is the mean and $\frac{s}{\sqrt{n}}$ is the Standard Error of the Mean
- z_c is the z-score (the number of SDs for a given CI)

Continuous Variables

- Commonly used confidence intervals include 90%, 95% and 99%
- The z-score is then calculated by looking at the area under the normal distribution curve at the specified confidence level
- Let us assume a confidence interval of 95% then we know that there is a remaining 5% of values that lie outside.
- To be exact, we know that 2.5% of values at each end of the tail of the normal distribution falls outside the 95% confidence interval

Continuous Variables



Continuous Variables

- Looking up the z-score for an area of 2.5% or 0.025 we get a corresponding value of 1.96
- We can now calculate the confidence interval using the z-score, the standard deviation and the sample size

$$\bar{x} \pm z_c \frac{s}{\sqrt{n}} \qquad 201 \pm 1.96 \frac{12}{\sqrt{500}} = 201 \pm 1.05$$

- This states that at a 95% confidence level the confidence interval is from 199.95 to 202.05

Smaller samples

- If the sample size is in fact smaller than 30 then we cannot assume a normal distribution anymore.
- We therefore need to use a Student's t-distribution which has fatter tails
- The distribution will result in larger confidence intervals compared to normally distributed data

What does the CI mean

- Saying you are 95% confident means that if you could take random samples repeatedly from this population and compare the confidence interval from each sample then in the long run 95% of these intervals would contain the mean.

Hypothesis Tests

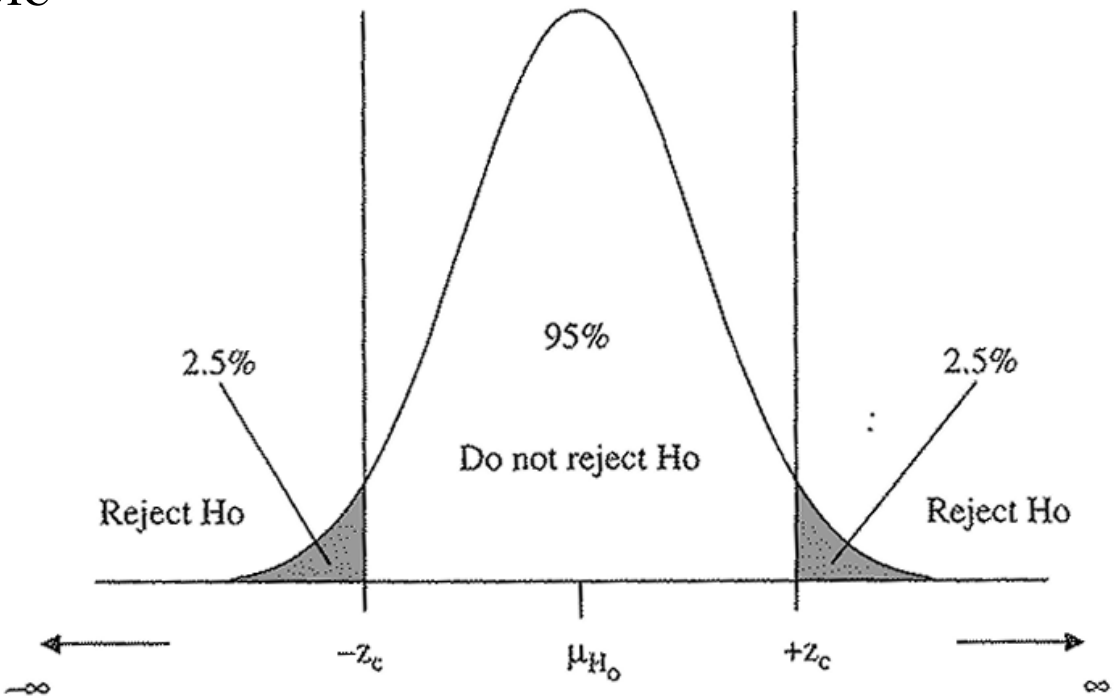
- The hypothesis test determines whether you have enough data to reject or not reject the 'Null Hypothesis'
  - Null Hypothesis (H_0): e.g. The average time to process a passport is 12 days.
  - Alternative Hypothesis (H_a): This is the conclusion that we would be interested in reaching if the null hypothesis is rejected. E.g. The average time to process a passport is not equal to 12 days.

Hypothesis Assessment

- Before the hypothesis test is performed we need to set a value at which H_0 should be rejected.
- As we are dealing with a sample the Hypothesis test may be incorrect
- This error can be minimised by choosing an appropriate confidence level
- This confidence level is usually described by the term α which is 100 minus the confidence percentage level divided by 100.
- E.g. A 95% confidence level has $\alpha = 0.05$ and a 99% confidence level has $\alpha = 0.01$

Hypothesis Assessment

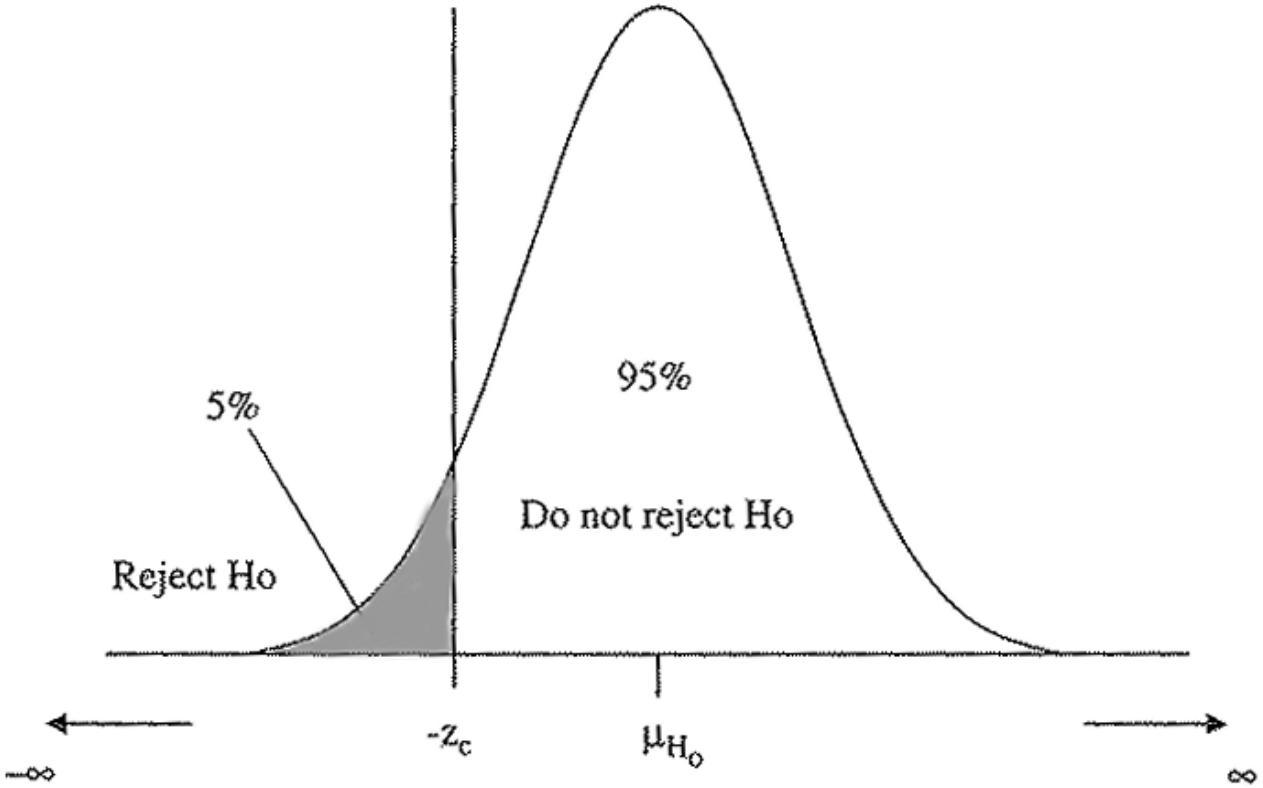
- First we calculate the statistic of interest from the sample
- The hypothesis test will then look at the difference between the value claimed in the hypothesis statements and the calculated sample



Hypothesis Assessment

- The graph on the previous slide indicated a two tailed test meaning that we reject if the value is less than or greater than.
- If the alternative hypothesis states only one of these conditions (e.g. Ha: $\mu < \mu_{H_0}$ and $\alpha = 0.05$) then we would reject the null hypothesis if the hypothesis test result has a z-score to the left of the critical value of z.
  - Since this is a one-tailed test the single area shaded should be equal to 5%

Hypothesis Assessment



Hypothesis Steps

1. Develop the hypothesis
2. Generate a sample
3. Calculate the summary statistics
4. Choose the statistical test
5. Calculate the test statistic
6. Derive the sampling distribution or find the correct table
7. Choose your cut-off value
8. Reject or fail to reject the null hypothesis
9. Draw a conclusion

Calculating p-values

- A hypothesis test is usually converted into a p-value.
- The p-value is the probability of getting the recorded value or a more extreme value
- It is measured of the likelihood of the result given the null hypothesis is true or the statistical significance of the claim
- Where the alternative hypothesis is not equal then the area under the curve value is doubled
- p-values range from 0 to 1
- Where the p-value is less than α the null hypothesis is rejected. If the p-value is greater than α the null hypothesis is not rejected

Exercise

- See exercises 1 and 2