[WEEK 2] Inferential Statistics

CS5701 - Quantitative Data Analysis

Dr Isabel Sassoon

Department of Computer Science
Brunel University London
isabel.sassoon AT brunel.ac.uk

28th September 2022

Brunel University London

Welcome to QDA – House keeping rules

- I will start at 10:05 prompt so please be in the Lecture theatre promptly
- The Lecture is being **recorded** - the recordings will be shared in Brightspace but it can take a few hours to make them available
- There will be time for questions, but we run out of time ask in the labs or post on the discussion forum on Brightspace

Learning Outcomes

This week after the lecture, the lab and the independent practice you should be:

- able to make a link between samples and the population
- able to interpret a histogram and density plot
- aware of different probability distributions
- able to use R to calculate **confidence intervals** for mean and proportion

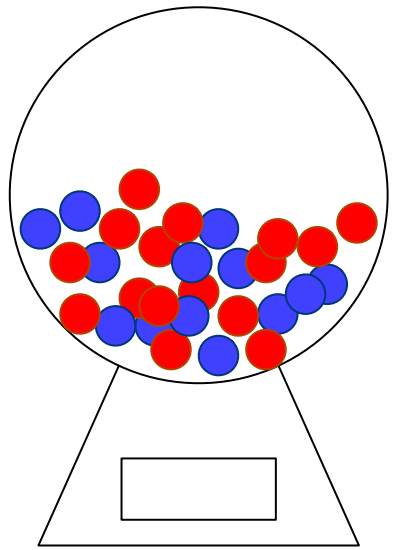
In the lab we will work on how to do these in RStudio

From Descriptive to Inferential Statistics

- In the previous lecture we looked at different data types and how to describe their contents, numerically and graphically
- Now we move to **using statistics to make inferences from sampled or available data to the population of interest**
- for example, given a sample mean....what does this tell us about the population mean?
- OR if we know a proportion from a sample... what does it tell us about the population proportion?

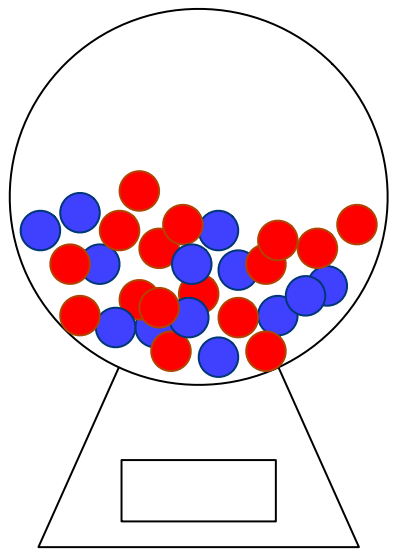
Gumball machine

- This is a Gumball machine with two flavours of sweets (Blue and Red)
- There are thousands of sweets in the machine and we cannot see inside
- Every time a coin is inserted 10 sweets come out



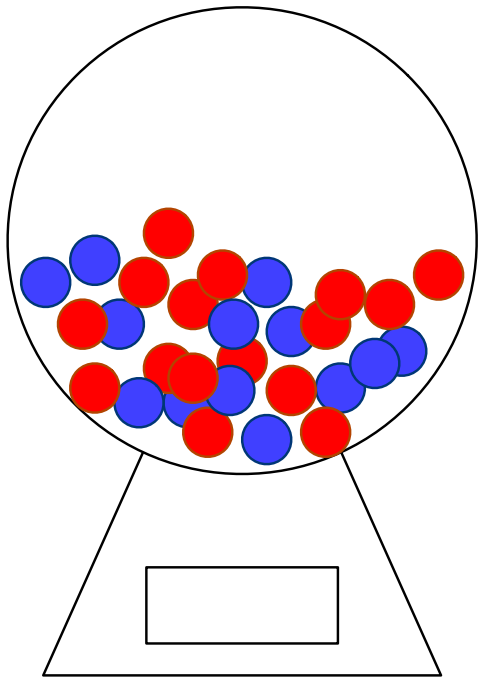
Gumball machine

- Every time a coin is inserted 10 sweets come out
- How many of the 10 will be red?
- What is the proportion of red sweets in the Gumball machine?
- How can we find out?



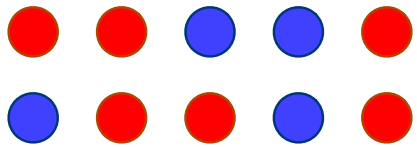
Gumball Machine - 1 sample of 10 sweets

What is the estimate for the proportion of red sweets from this sample?



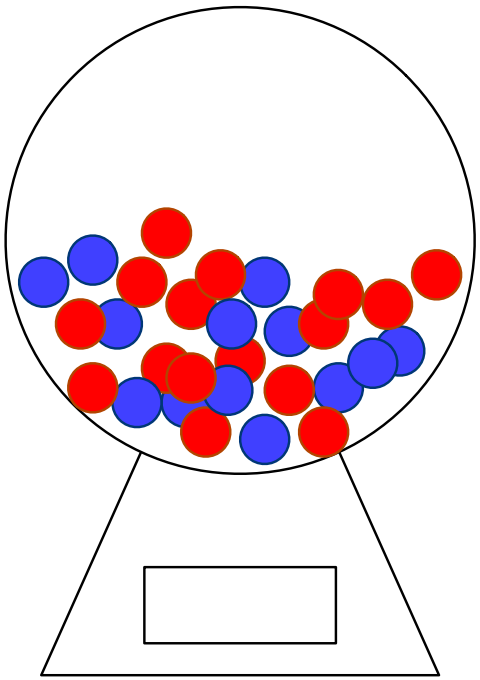
1 coin buys 10 sweets

There are 6 red sweets



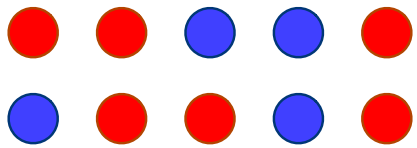
Gumball Machine - 1 sample of 10 sweets

What is the estimate for the proportion of red sweets from this sample?



1 coin buys 10 sweets

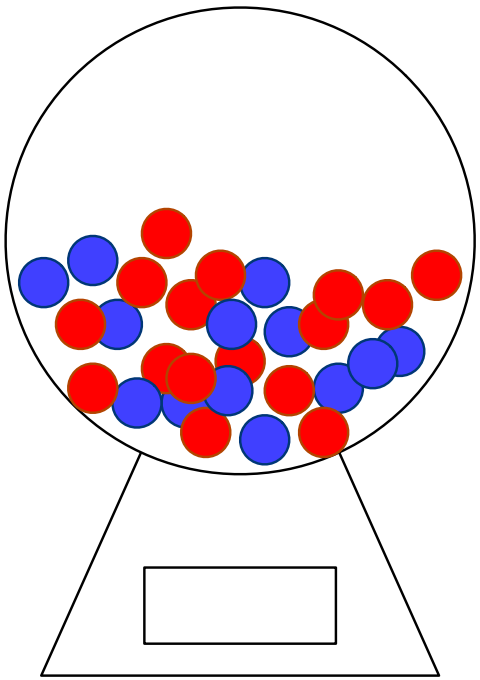
There are 6 red sweets



The proportion in the sample is $\frac{6}{10} = 0.6$

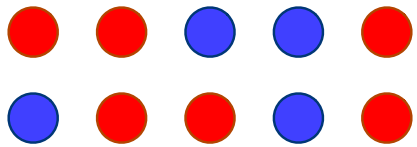
Gumball Machine - 1 sample of 10 sweets

What is the estimate for the proportion of red sweets from this sample?



1 coin buys 10 sweets

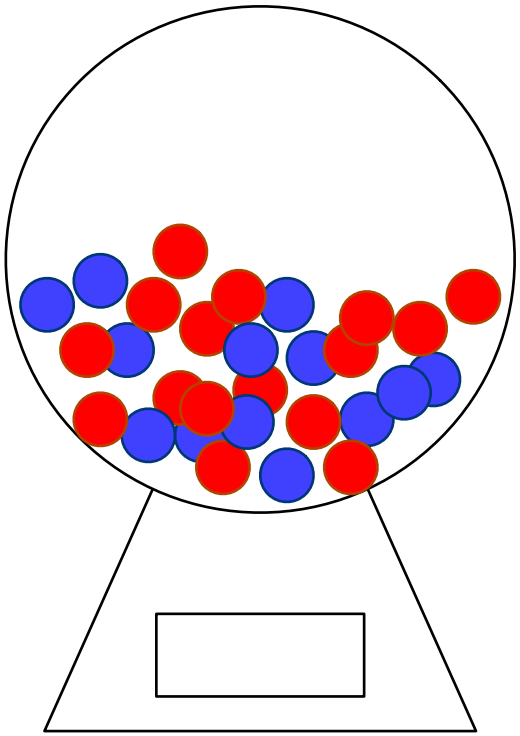
There are 6 red sweets



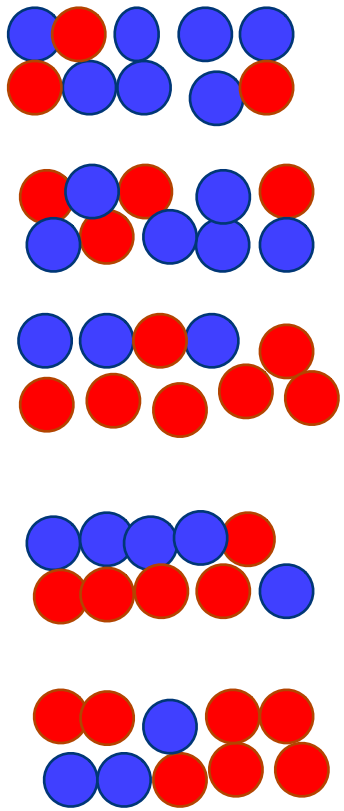
Is this a good guess for the proportion of red sweets in the whole machine?

Gumball Machine - 5 samples of 10 sweets

Let's take some more samples....
What is the estimate for the proportion of red sweets from these samples?

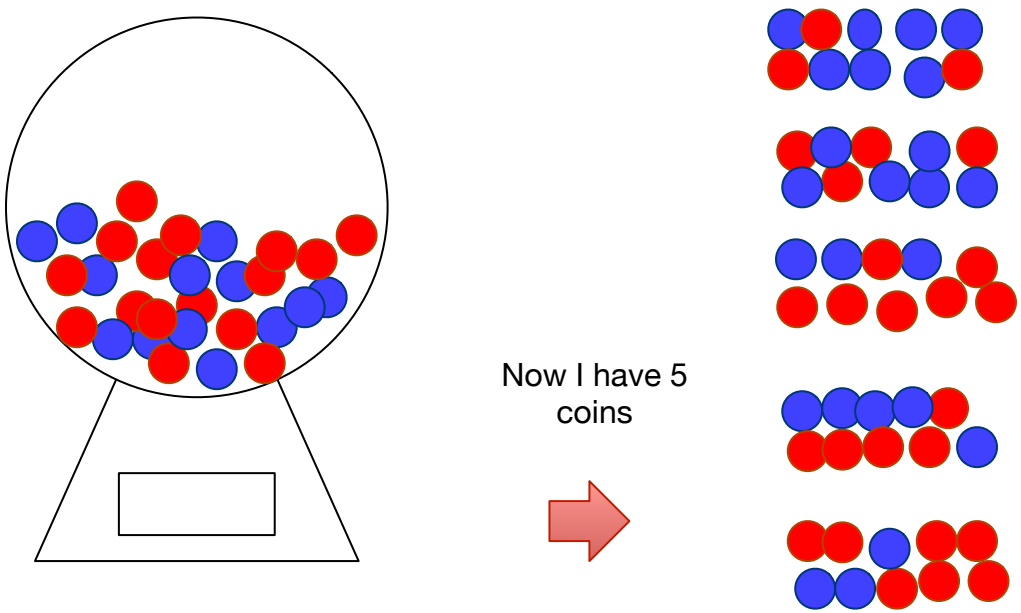


Now I have 5 coins



Gumball Machine - 5 samples of 10 sweets

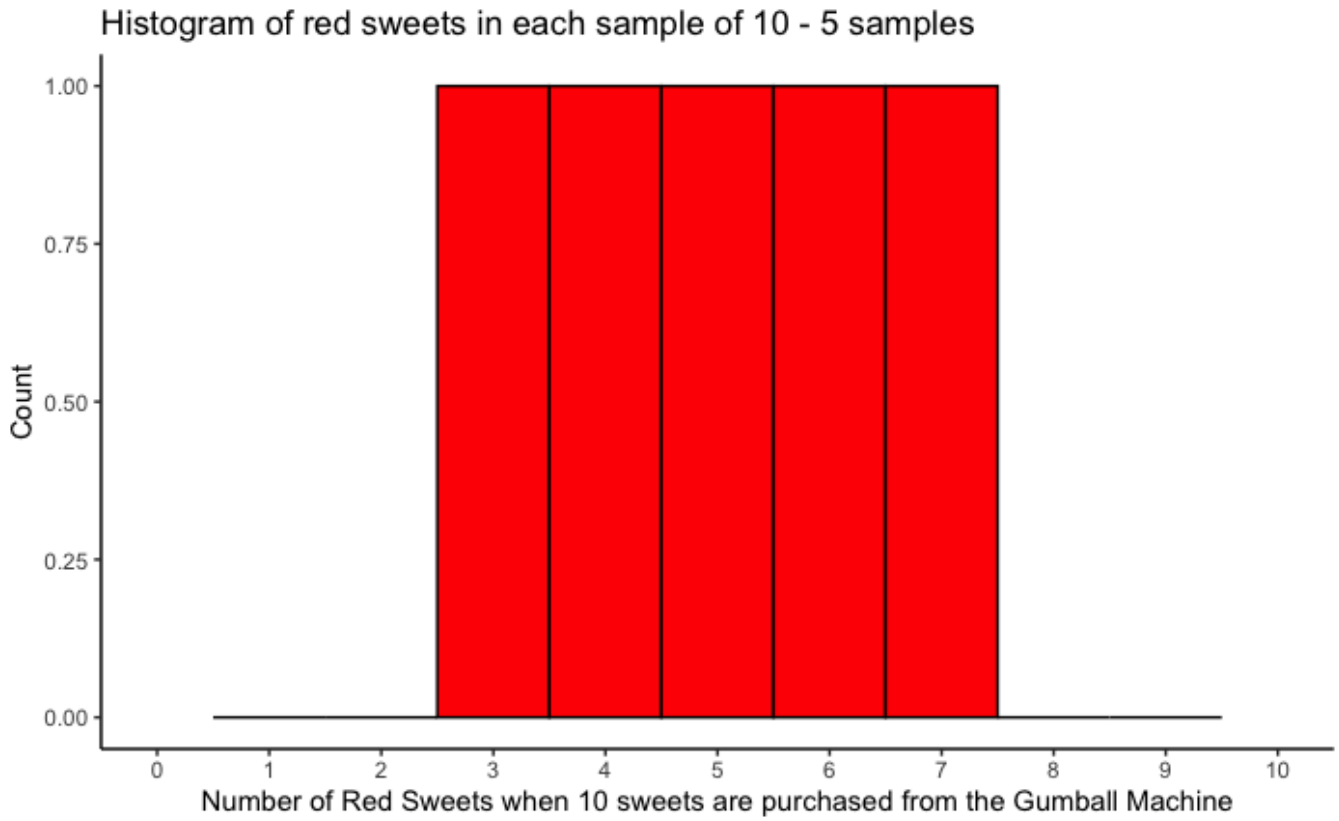
What is the estimate for the proportion of red sweets from these samples?



- These 5 samples have a different number of red sweets each: $3, 4, 7, 5, 6$
- so the proportions in each of these is $0.3, 0.4, 0.7, 0.5, 0.6$
- if we look at the mean for these its 0.5

Gumball Machine - 5 samples of 10 sweets - Histogram

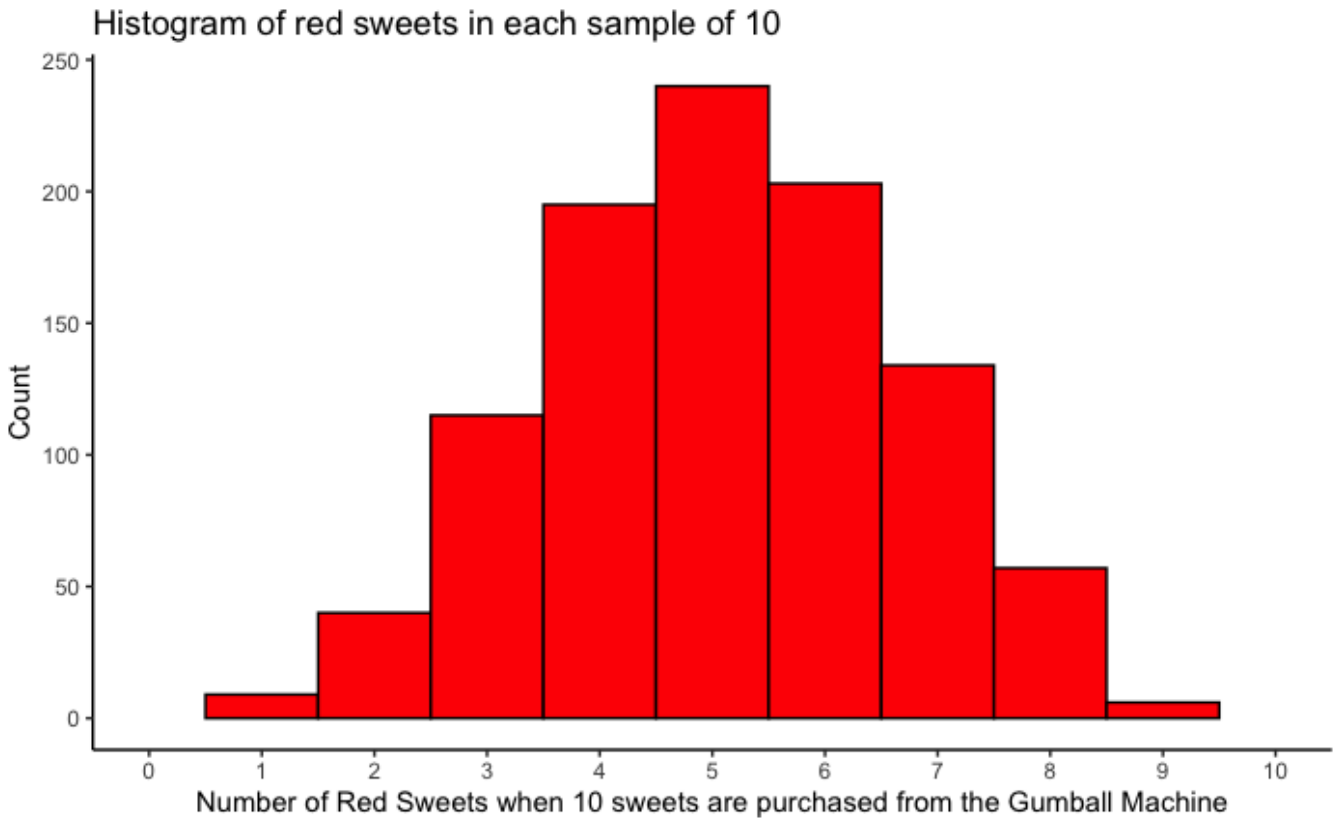
We can explore the distribution of number of red sweets in each sample of 10 using a histogram:



What do we see? How can this be improved?

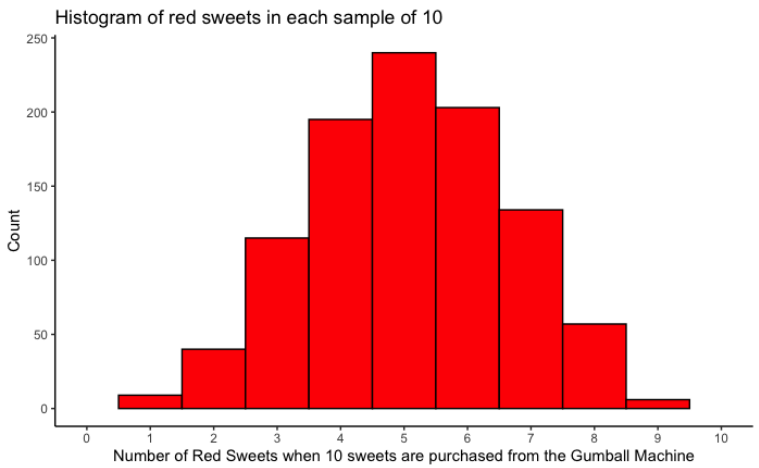
Gumball Machine - 1000 samples of 10 sweets - Histogram

Now we have 1000 samples of 10 sweets from this gumball machine:



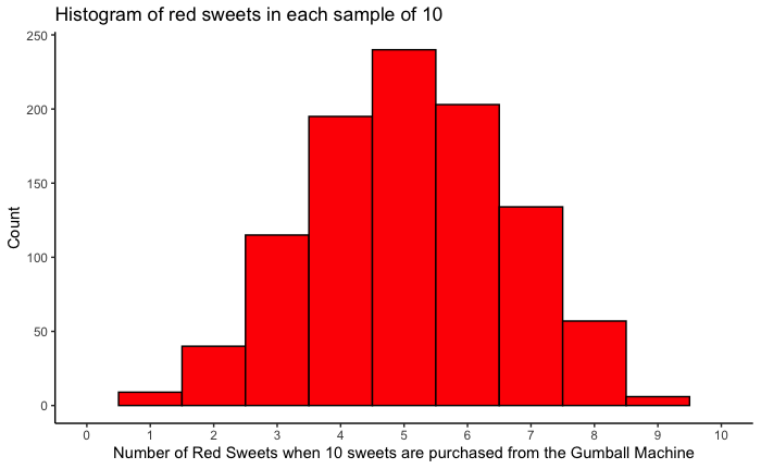
Gumball machine - what is the proportion of red sweets?

- The histogram illustrates the distribution of the possible values
- Even if there are exactly 50% (or a proportion 0.5) red sweets in the Gumball machine, every time we sample 10 we will get a different no. of reds
- As we take more and more samples we can see which values are more likely than others



Gumball machine - what is the proportion of red sweets?

- From the histogram we can see that 5 out of 10 red sweets is the most frequent outcome
- The mean proportion of red sweets computed from the 1000 samples is 0.500124
- We can even read from this histogram that only about 25 samples resulted in 1 red sweet only or 9 red sweets
- In **97.5% of the samples the number of red sweets was between 2 and 8**



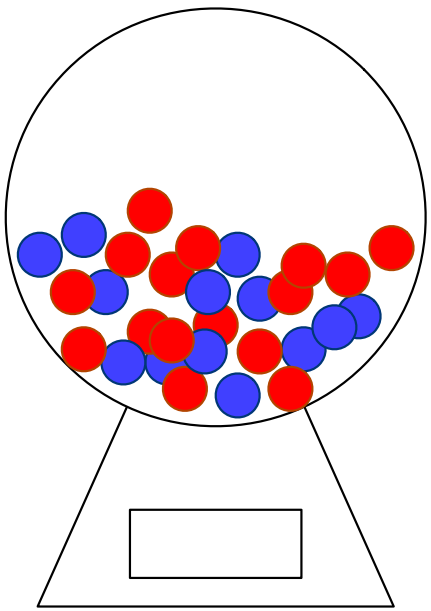
Brunel University London

Confidence Intervals

- A **Confidence Interval** shows us the likely range in which the mean (or the proportion) would fall if the sampling exercise were repeated.
- The more confident you want to be the wider the interval will be
- Confidence intervals can be produced at different level of confidence - typically 95% confidence interval is used

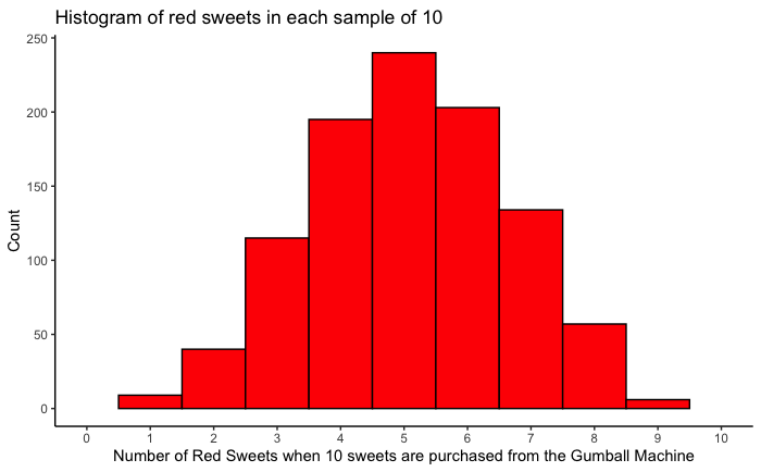
Confidence Intervals for the Gumball Machine

- We have 1000 samples and have computed the proportion of red sweets for each one
- Our estimate for the proportion is 0.49 (the mean of these 1000 proportions)
- Confidence intervals are the likely range in which the TRUE proportion would fall



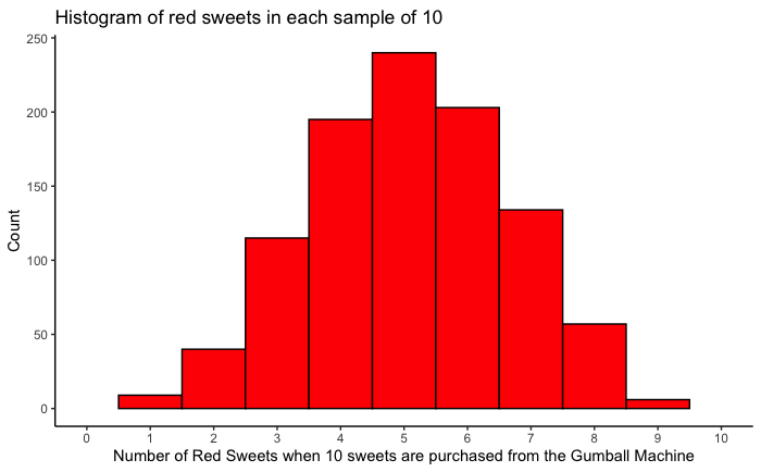
Gumball machine - what is the proportion of red sweets?

- From the histogram we can see that 5 out of 10 red sweets is the most frequent outcome
- In 97.5% of the samples the number of red sweets was between 2 and 8
- **So the intuition from this is that, given the probability distribution from the sample, we are 97.5% confident that the real value for the proportion of red sweets is between 0.2 and 0.8.**



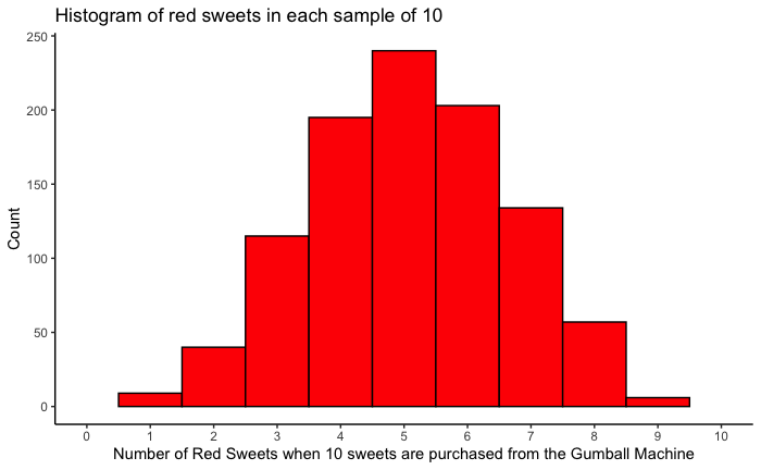
Gumball machine - what is the proportion of red sweets?

- In 97.5% of the samples the number of red sweets was between 2 and 8
- So the intuition from this is that, given the probability distribution from the sample, we are 97.5% confident that the real value for the proportion of red sweets is between 0.2 and 0.8.
- Following the same how confident could we be that the real value for the proportion of red sweets is between 0.4 and 0.6?



Gumball machine - what is the proportion of red sweets?

- Following the same how confident could we be that the real value for the proportion of red sweets is between 0.4 and 0.6?
- Approximately **65% of the samples resulted in a proportion between 0.4 and 0.6, se we are 65% confident that the real value for the proportion of red sweets is between 0.4 and 0.6.**



Confidence Intervals - more formal

If we want to compute a **confidence interval** for an unknown parameter θ (for example this could be the population mean)

Definition

Let $X = (X_1, \ldots, X_n)$ be a random sample from $f(x, \theta)$. Let $A(X)$ and $B(X)$ be two statistics. Then the interval $[A(X), B(X)]$ is called the $100(1-\alpha)\%$ confidence interval for θ if:

$$P(A(X) \leq \theta \leq B(X)) = 1 - \alpha$$

Intuitively the probability that θ is between those two values is $1 - \alpha$

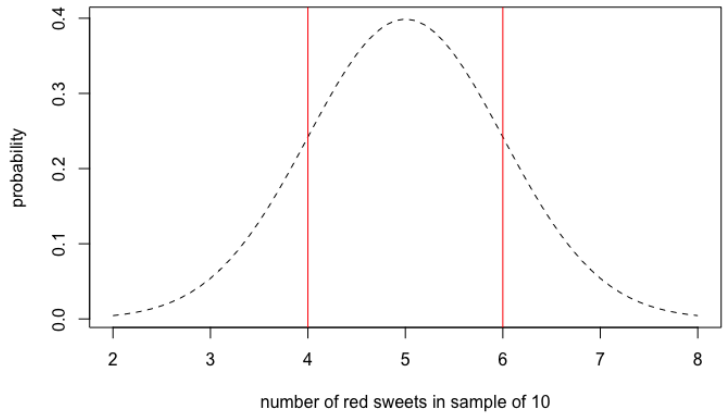
Typical values for α are 0.05, 0.1, 0.01. 0.05 is equivalent to 95% confidence level.

Confidence Intervals for the proportion

- To compute confidence intervals for **proportions**
- Assuming that π is the population proportion, p the sample population, a sample size of n
- X_i is a random variable such that $X_i = 1$ if the answer to the question is "yes", or $X_i = 0$ if the answer is "no"
- The estimated proportion p is therefore $\frac{\Sigma_{i=1}^{n} X_i}{n}$
- Then the confidence interval is $p \pm z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$
- Where z is the critical value from the Normal distribution
- This is only valid with large sample sizes, a good heuristic $np > 5$

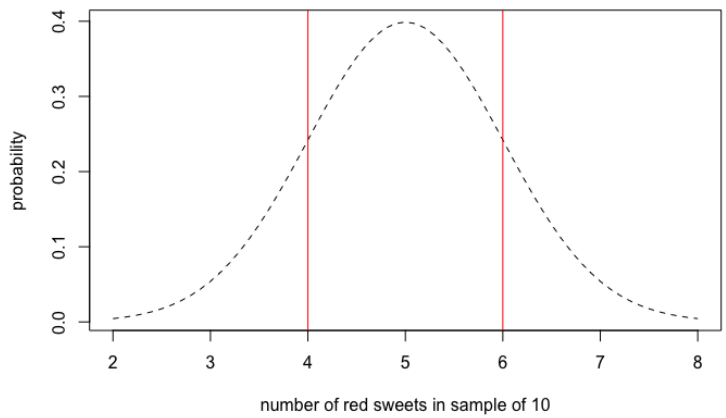
Confidence Interval from the Gumball example

- This is a density plot
- It is a smoothed version of the histogram - using a normal distribution
- The area under the curve sums up to 1



Confidence Interval from the Gumball example

- In the density plot $\rightarrow$
- Approximately 65% of the samples resulted in a proportion between 0.4 and 0.6, se we are 65% confident that the real value for the proportion of red sweets is between 0.4 and 0.6.
- Intuitively that 65% is equivalent to the area under the curve between the red lines.



Example - Gumball machine

- We have our gumball machine and have taken 100 samples of 10 sweets (*)
- We counted the number of red sweets in each of the 100 samples, computed the proportion in each sample and then the mean of those proportions
- This mean proportion is 0.45
- Lets find a 95% and a 90% confidence interval for the true proportion of red sweets in the gumball machine

(*)This is a different sample of 100

Example - Gumball machine

We have our gumball machine and have taken 100 samples of 10 sweets, counted the number of red sweets in each of the 100 samples, computed the proportion in each sample and then the mean of those proportions. This mean proportion is 0.45.

Lets find a 95% and a 90% confidence interval for the true proportion of red sweets in the gumball machine.

- $p = 0.45$, $n = 100$ and the values for $Z_{0.025} = 1.96$
- Recall the equation for the CI for proportions:

$$p \pm z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$$

Example - Gumball machine

This can be computed in R using:

- `prop.test(45,100, conf.level=0.9)`
- OR
- `binom.test`

...and the results are:

- 95% percent confidence interval: (0.350.55)
- 90% percent confidence interval: (0.360.53)

Effect on CI size

What happens to the Confidence Intervals width when:

- if we have more sample data or a larger sample size?
- if we have a higher confidence level?

GO TO: pollev.com/isabelsassoon

Confidence Intervals for the mean

If we want to compute the confidence interval for the mean μ given a random sample $X_1, \ldots, X_n$ we need to take the following steps:

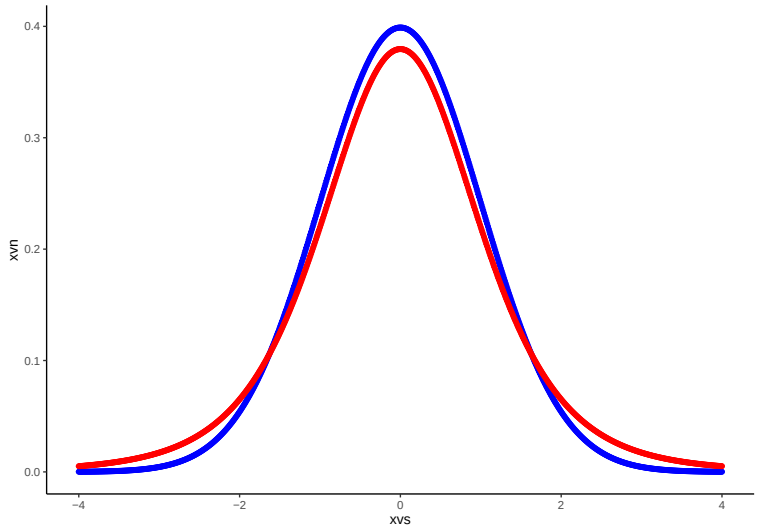
- Compute the sample mean $\bar{x}$ and the sample standard deviation s
- Decide what confidence interval level to use (e.g. 95%, 90% or other) and find the Z value for that level (for 95% $Z = 1.96$)
- Use this formula to compute the interval:

$$(\bar{x} - Z \times \frac{s}{\sqrt{n}}, \bar{x} + Z \times \frac{s}{\sqrt{n}})$$

In R use `qnorm(0.975)=1.96` to find the Z value

to Z or T?

- **Student's t** distribution is used instead of the normal distribution when sample sizes are small ($n < 30$)
- **Student's t** produces bigger intervals
- t and Z are similar already for $df = 5$ (see the plot)



The confidence Interval using t:

$$(\bar{x} - t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}, \bar{x} + t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}})$$

Brunel University London

Computing confidence intervals for the mean - example

Given a sample of 30 observations with $\bar{x} = 30.969$, $s = 18.36$ we can now compute the confidence intervals using Z and t

- Confidence interval is 95% and therefore $Z = 1.96$
- The confidence interval is:
  $(30.969 - 1.96 \times \frac{18.36}{\sqrt{30}}, 30.969 + 1.96 \times \frac{18.36}{\sqrt{30}}) = (24.4, 37.5)$
- If we used t with 30 df and the same confidence level the interval would be: (24.1, 37.8)

to find the value for t in R use `qt(0.975,29)=2.05` for z use `qnorm(0.975)=1.96`

Brunel University London

example - shearing strength

The shearing strength (in tonnes) of a steel rivet of a certain type was measured for a random sample of 12 rivets. These were the measurements:
$4.01, 4.28, 4.43, 4.04, 3.74, 4.43, 3.89, 4.73, 3.68, 4.55, 4.10, 3.77$
Given these measurements we can:

- Estimate the mean shearing strength
- Estimate the standard deviation for this sample mean
- Then we can build a confidence interval for the mean shearing strength

To compute the sd from the sample:

$$\bar{s} = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

Brunel University London

example - shearing strength contd.

- The mean shearing strength μ can be estimated by computing the sample mean $\bar{x} = 4.1375$
- The populations s.d. σ can be estimated using the s.d. of the sample $s = 0.344$
- Given these we can compute the confidence interval - but should we use t or Z?

example - shearing strength contd.

- The mean shearing strength μ can be estimated by computing the sample mean $\bar{x} = 4.1375$
- The populations s.d. σ can be estimated using the s.d. of the sample $s = 0.344$
- Given the sample size t is the one to use

Recall:

$$(\bar{x} - t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}, \bar{x} + t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}})$$

Brunel University London

example - shearing strength contd.

- Given the small sample size and the need to estimate the s.d from the sample too - we should use t
- The $t_{0.025,11} = 2.20$, recall if we want 95% confidence intervals then this equates to using 0.025 and $df = 11$
- The confidence interval is therefore:

$$4.138 \pm 2.20 \times \frac{0.344}{\sqrt{12}} = 4.138 \pm 0.2178$$

  which is $(3.92, 4.36)$

This can be computed in R using: `t.test(shear, conf.level=0.95)`

Summary

In this lecture we have:

- to made a link between samples and the population
- interpreted a histogram
- computed confidence intervals for mean and proportion
- IN THE LAB you will use R Studio to calculate **confidence intervals** for mean and proportion