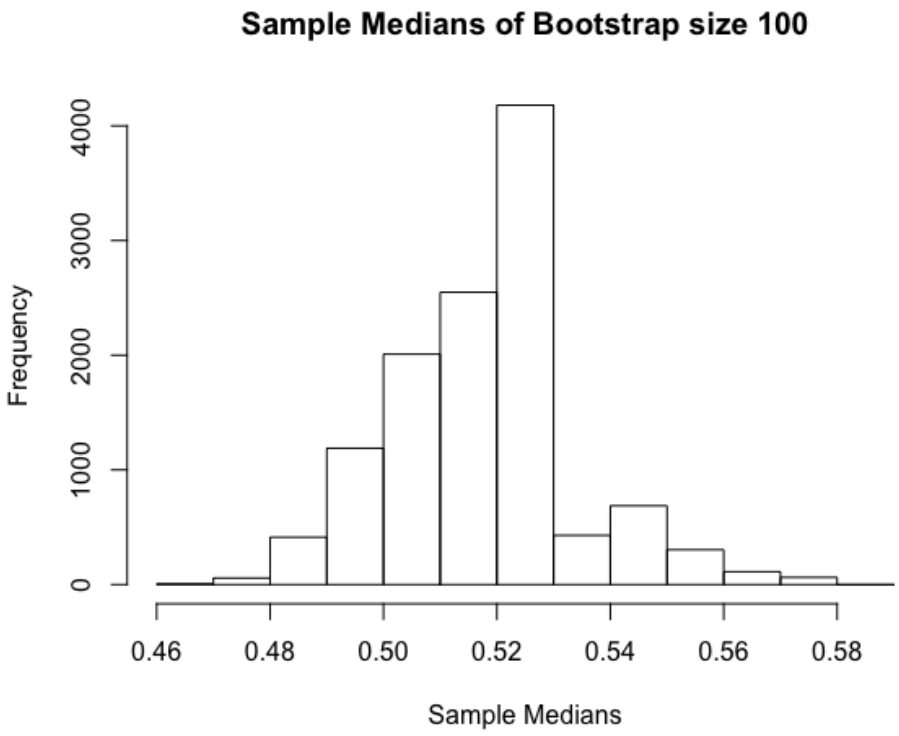
1. **6.10 Investigating the construction of bootstrap confidence intervals** The UC Irvine Machine Learning data repository hosts a dataset giving various measurements of abalone at https://archive.ics.uci.edu/ml/datasets/Abalone. This data comes from an original study by W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn and W. B. Ford, called "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288) (1994). The data was donated by S. Waugh. There are 4177 records. We will use the Length measurement. We will assume that the 4177 records is the entire population. Compute the population median.

    a. **(a)** Draw 10,000 samples of 100 records at random with replacement. Use each sample to produce a bootstrap estimate of a centered 90% confidence interval for the population median. For what fraction of samples does the true population median lie inside the interval?

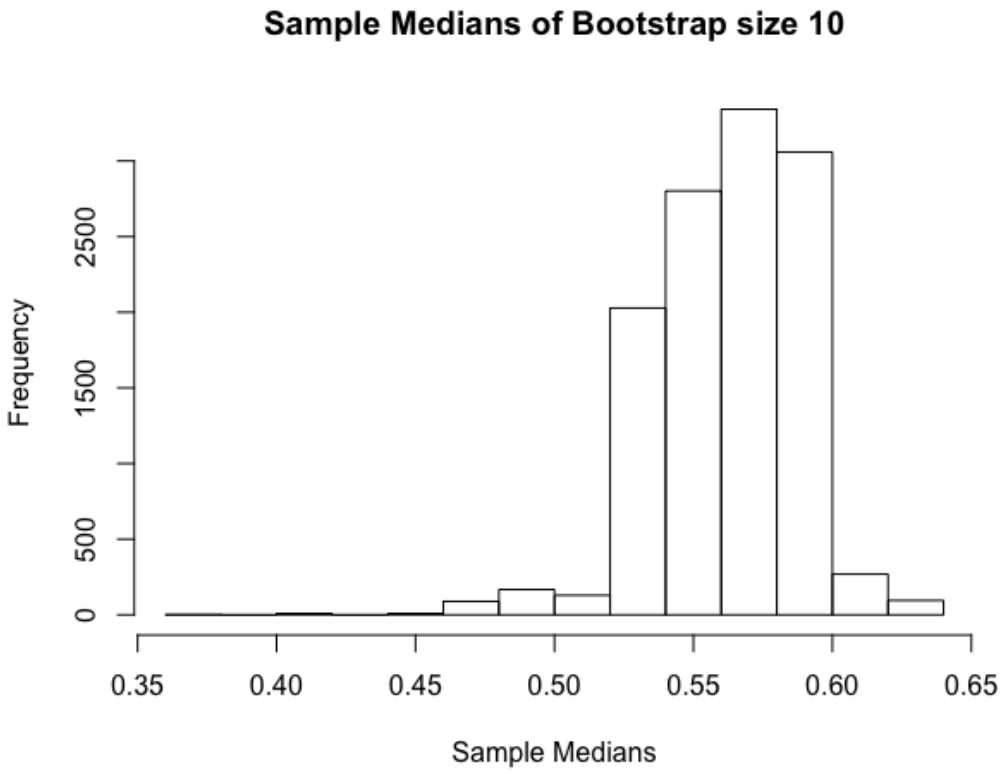
       Population median = 0.545. Out of the 10,000 bootstrap simulations 8789 of them had the population mean within the %90 confidence interval. Meaning that 0.8789 of the samples had the true population median within them.



    b. **(c)** Draw 10,000 samples of 10 records at random with replacement. Use each sample to produce a bootstrap estimate of a centered 90% confidence interval

for the population median. For what fraction of samples does the true population median lie inside the interval?

Population median is 0.545 same as the previous question. Out of the 10,000 bootstrap simulations, 8687 of them included the population median within the %90 confidence interval. So, 0.8687 of the samples included the population mean within their confidence interval.



c. **(d)** What conclusion do you draw?

As the record numbers in samples increase, there is an increase on the reliability of the bootstrap. 90% should be the fraction of the bootstrap simulation that has the true median within its %90 confidence interval, and 0.8789 is closer to 90% than 0.8687.

2. **7.1** In 1998, the average height of an adult male in South Africa was estimated to be 169 cm. Assume that this estimate is exact; assume also that the population standard deviation is 10 cm. What fraction of samples consisting of 50 adult males from South Africa (selected uniformly at random, and with replacement) will have average height greater than 200 cm?

   Standard error will be equal to standard deviation divided by the square root of the sample size: $stderr = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{50}} = 1.41421356237$. Now that we have the standard error and since the population mean is given as 169, we can write the sample mean as a t-distributed random variable:

   $$T = \frac{SampleMean - PopMean}{stderr} = \frac{SampleMean - 169}{1.41421356237}$$

   With N-1 = 49 degrees of freedom.
   Since there are enough degrees of freedom, we can now form an integral with the standard normal probability distribution using the t-distributed random variable and set $\frac{200-169}{1.41421356237} = 21.9203102168$ as the lower bound. And set the upper bound to be infinity.

   $$\int_{21.92}^{\infty} \frac{1}{\sqrt{2*\pi}} e^{\left(\frac{-u^2}{2}\right)} du = 8.31601*10^{-107}$$

   According to my integral calculator $8.31601*10^{-107}$ fraction of the samples will have an average height greater than 200. And this, conceptually makes sense since %99.7 of the data is within 3 standard deviations of the mean while 200 is 6 standard deviations away from the mean, meaning that having an average height of >200 is almost impossible.

3. **7.2** Assume the average weight of an adult male short-hair house cat is 5 kg, and the standard deviation is 0.7 kg (these numbers are reasonable, but there's quite a lively fight between cat fanciers about the true numbers).
    a. **(a)** What fraction of samples consisting of 30 adult male short-hair house cats (selected uniformly at random, and with replacement) will have average weight less than 4 kg?

        We know the population mean to be 5 and then we can find the standard error by dividing the population standard deviation by the square root of sample size.

        $stderr = \frac{\sigma}{\sqrt{N}} = \frac{0.7}{\sqrt{30}} = 0.12998673672$ and now we can set up a t-distribution for the mean of the sample as a random variable:

        $$T = \frac{SampleMean - PopMean}{stderr} = \frac{SampleMean - 5}{0.12998673672}$$

        For sample mean = 4, T will be -7.69309258185. Since we have enough degrees of freedom, we can use the standard normal distribution. which will be the upper bound of the integral since we want the avg. weight < 4, and the lower bound will be negative infinity.

        $$\int_{-\infty}^{-7.69309258185} \frac{1}{\sqrt{2*\pi}} e^{\left(\frac{-u^2}{2}\right)} du = 7.18103 * 10^{-15}$$

        So, almost none of the samples will have an average weight less than 4 kg.

    b. **(b)** What fraction of samples consisting of 300 adult male short-hair house cats (selected uniformly at random, and with replacement) will have average weight less than 4 kg?

        Now our standard error will be different since N is 300 now,

        $stderr = \frac{\sigma}{\sqrt{N}} = \frac{0.7}{\sqrt{300}} = 0.04041451884$

        Now we can set up the t-distributed random variable as:

        $$T = \frac{SampleMean - PopMean}{stderr} = \frac{SampleMean - 5}{0.04041451884}$$

        For sample median to be 4, the upper bound of the integral will be equal to -24.7435829673. Then the lower bound will be negative infinity since we are interested in the cumulative probability.

        $$\int_{-\infty}^{-24.7435829673} \frac{1}{\sqrt{2*\pi}} e^{\left(\frac{-u^2}{2}\right)} du = 1.8174 * 10^{-135}$$

        Meaning that even a smaller fraction of different samples will have an average weight below 4.

    c. **(c)** Why are these numbers different?

This is because standard error depends on the sample size and standard error is used to calculate the upper bound of the integral, meaning that we have to get a different result for both.

4. **7.4 How big are Parktown Prawns?** The Parktown prawn is an impressively repellent large insect, common in Johannesburg (look them up on the Web). I claim that their average length is 10 cm. You collect 100 Parktown prawns (this will take about 10 mins, in the right places in Johannesburg; more difficult from the US). The mean length of these prawns is 7 cm. The standard deviation is 1 cm. Assess the evidence against my claim.

   This can be written as t-distribution of 99 degrees of freedom. Since 99 > 30 we can just use the normal probability distribution for our probability density function. The standard error is going to be the unbiased standard deviation divided by the square root of the sample size.

$$stderr = \frac{stdunbiased}{\sqrt{N}} = \frac{samplestd}{\sqrt{N}} = \frac{1}{\sqrt{100}} = 0.1$$

   Assuming popmean = 10, then the random variable T with t-distribution looks like:

$$T = \frac{SampleMean - PopMean}{stderr} = \frac{7-10}{0.1} = -30$$

   The integral would use the absolute value of T for the upper bound and the negative of the absolute value of T for the lower bound. Then the p-value would be:

$$p = (1 - f) = 1 - \int_{-30}^{30} \frac{1}{\sqrt{2*\pi}} e^{\left(\frac{-u^2}{2}\right)} du$$

$p \cong 0$

   A smaller p-value means that the hypothesis is more likely to be wrong, therefore I **reject the claim** since the p value is approximately zero.

5. **7.8 Are boys and girls equiprobable?** In Carcelle-le-Grignon at the end of the eighteenth century, there were 2009 births. There were 983 boys and 1026 girls. You can regard this as a fair random sample (with replacement, though try not to think too hard about what that means) of births. Assess the evidence against the hypothesis that a boy is born with probability exactly 0.5.

   Let's assume that each boy is mapped to 1 while each girl is mapped to 0, then the sample mean would become 0.48929815828 and the sample standard deviation becomes 0.5000. We remove the bias from the standard deviation by multiplying it with sqrt(N/N-1):

$$stderr = \frac{stdunbiased}{\sqrt{N}} = \frac{\sqrt{\frac{N}{N-1}} * samplestd}{\sqrt{N}} = \frac{0.5000}{\sqrt{2009}} = 0.01115548992$$

   Now, we assume that the population mean is 0.5 like in the hypothesis and we write the t-distributed random variable T as:

$$T = \frac{SampleMean - PopMean}{stderr} = \frac{0.48929815828 - 0.5}{0.01115548992} = -0.95933408543$$

   Since the degrees of freedom is 2008, we can just use the standard normal probability distribution for our probability distribution function. Now we define the p-value as:

$$p = (1 - f) = 1 - \int_{-0.95933408543}^{0.95933408543} \frac{1}{\sqrt{2 * \pi}} e^{\left(\frac{-u^2}{2}\right)} du$$

$$p = 0.33739$$

   The p value is relatively big, meaning that there isn't enough evidence to reject the claim, so **I do not reject the claim.**

6. **7.10** You can find a dataset giving income data for US citizens at the UC Irvine Machine Learning data archive, at http:// archive.ics.uci.edu/ml/datasets/Adult. Each item consists of a set of numeric and categorical features describing a person, together with whether their annual income is larger than or smaller than 50 K$.
    a. **(a)** Assess the evidence that income category is independent of gender.

        We set up a table that contains the frequency of earning above or below 50K $ for male and female, using 10^{th} and 15^{th} entries of the dataset.

	<=50K	>50K
Female	9592	1179
Male	15128	6662

        Then we use chisq.test() function, which is available in R. The p-value, then, is evaluated to be p < 2.2e-16, meaning that our p value is extremely small. Since the p-value is that small, we have enough evidence to reject the claim that income category is independent of gender. It's extremely unlikely that two categories are independent.

    b. **(b)** Assess the evidence that income category is independent of education level.

        Same as in part a, first we set up a frequency table using the built-in R table() function, the table consists of the 15^{th} and the 4^{th} columns of the dataset and looks like this:

	<=50K	>50K
10th	871	62
11th	1115	60
12th	400	33
1st-4th	162	6
5th-6th	317	16
7th-8th	606	40
9th	487	27
Assoc-acdm	802	265
Assoc-voc	1021	361
Bachelors	3134	2221
Doctorate	107	306
HS-grad	8826	1675
Masters	764	959
Preschool	51	0
Prof-school	153	423
Some-college	5904	1387

Using the R chi-square test function, the p-value was found to be smaller than 2.2e-16. Meaning that it's extremely unlikely for education level to be independent of income, therefore I reject the claim.