

# Assignment 1

Group 69

24.02.2023

The data set `birthweight.txt` contains the birthweights (in grams) of 188 newborn babies. Denote the underlying mean birthweight by  $\mu$ .

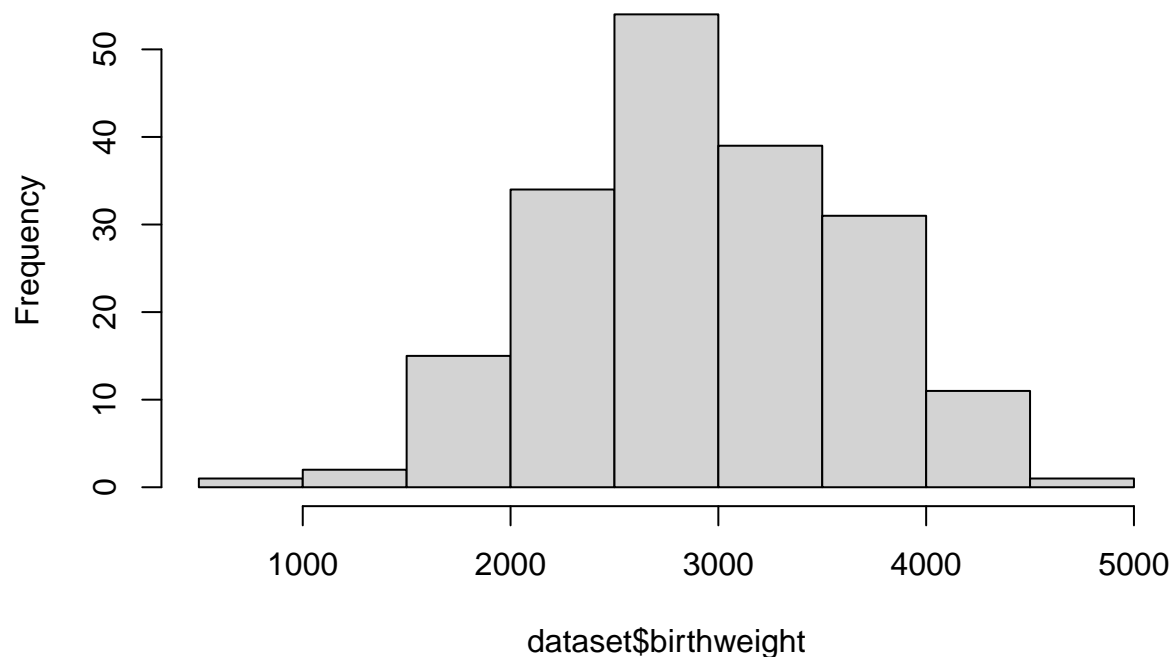
- a) Check normality of the data. Assuming normality (irrespective of your conclusion about normality), construct a bounded 96%-CI for  $\mu$ . Evaluate the sample size needed to provide that the length of the 96%-CI is at most 100. Compute a bootstrap 96%-CI for  $\mu$  and compare it to the above CI.

To start with we import the `birthweight.txt` dataset and store it to the respective variable.

To make an assumption about the distribution we display the data using the following function.

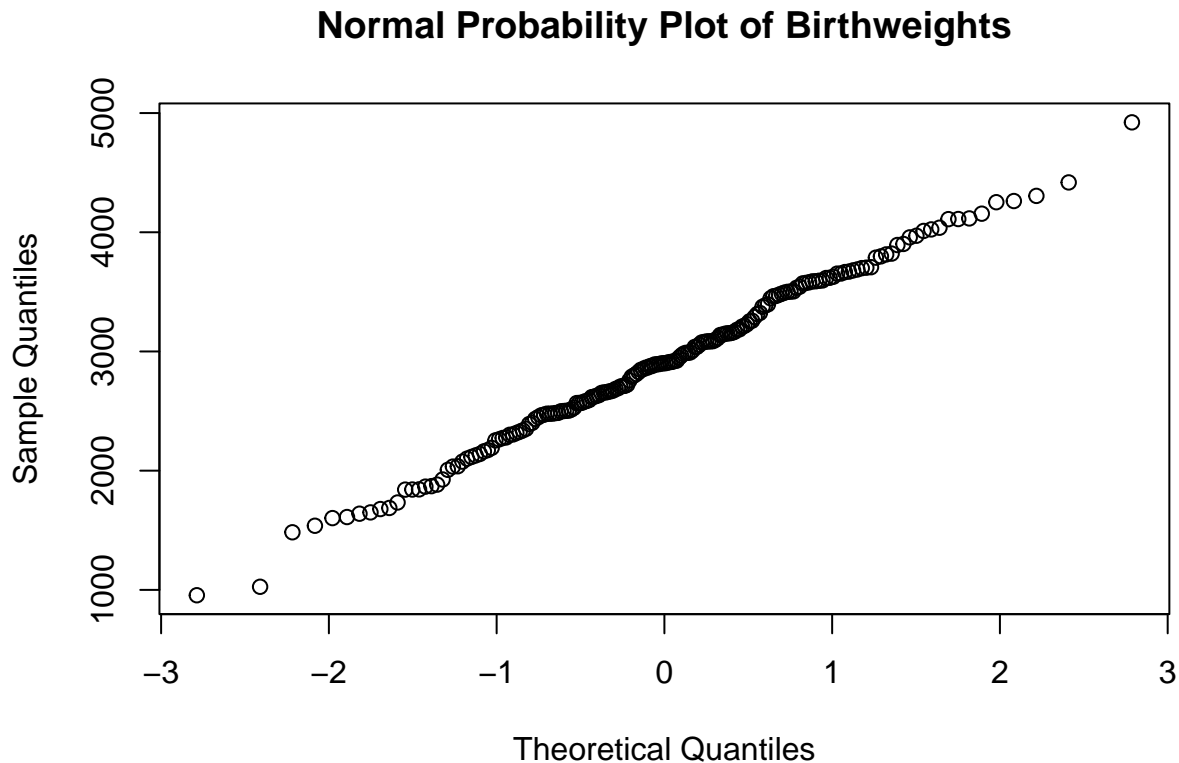
```
dataset = read.table("birthweight.txt", header = T)
hist(dataset$birthweight, main="Histogram of Birthweights")
```

## Histogram of Birthweights



After running the histogram we observe a bell-shaped form, which indicates that the data is normally distributed. Although we can clearly see the shape, the histogram alone isn't a good indicator of normally distributed data.

```
qqnorm(dataset$birthweight, main="Normal Probability Plot of Birthweights")
```



We run the `qqnorm` function to confirm the distribution. A nearly straight line indicates that the distribution is indeed normal.

To construct the 96%-CI for the  $\mu$  birthweight the following calculations are performed.

```
n <- length(dataset$birthweight)
se <- sd(dataset$birthweight) / sqrt(n)
t_star <- qt(0.98, df=n-1)
ci <- c(mean(dataset$birthweight) - t_star*se, mean(dataset$birthweight) + t_star*se)
```

To get the 96%-CI of the mean we find the sample size ( $n$ ), standard error ( $se$ ), and critical value of  $t$ -distribution ( $t_{star}$ ). We apply the formula  $\bar{\mu} \pm t * (se/\sqrt{n})$ .

To evaluate the sample size needed to provide that the length of the 96%-CI is at most 100.

```
t_star <- qt(0.98, df=1e6)
n <- ceiling(((t_star * sd(dataset$birthweight) / 50) ^ 2))
```

Here the  $t_{star}$  uses other  $df$  value that denotes an arbitrary large sample size that we need for our calculation.

To compute a bootstrap 96%-CI for  $\mu$  define a function to generate a bootstrap sample and compute the mean of the sample.

```
boot_mean <- function(data) {
  boot_sample <- sample(data, replace=TRUE, size=length(data))
  mean(boot_sample)
}
```

Generate 10000 bootstrap samples and store the means in a vector.

```
boot_means <- replicate(10000, boot_mean(dataset$birthweight))
```

Calculate the 2.5th and 97.5th percentiles of the bootstrapped means to obtain the 96% CI.

```
boot_ci <- quantile(boot_means, c(0.025, 0.975))
```

So the values of the bootstrap CI to the one obtained using the t-distribution could be compared.

```
ci
```

```
## [1] 2808.084 3018.501
```

```
boot_ci
```

```
##      2.5%      97.5%
## 2812.305 3010.506
```

- b) An expert claims that the mean birthweight is bigger than 2800 gram. Verify this claim by using a relevant t-test, explain the meaning of the CI in the R-output for this test. Also propose and perform a suitable sign tests for this problem.

To verify the claim a one-sample t-test could be performed with null hypothesis  $H_0: \mu = 2800$  and alternative hypothesis  $H_a: \mu > 2800$ . Here,  $\mu$  represents the true population mean birthweight.

```
t.test(dataset, mu=2800, alternative="greater")
```

```
##
## One Sample t-test
##
## data:  dataset
## t = 2.2271, df = 187, p-value = 0.01357
## alternative hypothesis: true mean is greater than 2800
## 95 percent confidence interval:
##  2829.202      Inf
## sample estimates:
## mean of x
##  2913.293
```

The output of this test includes the test statistic (t-value), the p-value, and a confidence interval for the population mean. The p-value represents the probability of observing a sample mean at least as extreme as the one we obtained, assuming that the null hypothesis is true.

The p-value is 0.01357 which is less than the significance level, so we reject the null hypothesis and conclude that there is strong evidence to suggest that the mean birthweight is indeed greater than 2800 gram. The confidence interval also supports this conclusion, as it does not include 2800 gram.

Alternatively a sign test can be performed to confirm the hypothesis.

```
n <- length(dataset$birthweight)
sign_test <- binom.test(sum(dataset$birthweight > 2800), n, p=0.5, alternative="greater")
sign_test

##
## Exact binomial test
##
## data:  sum(dataset$birthweight > 2800) and n
## number of successes = 107, number of trials = 188, p-value = 0.03399
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5065781 1.0000000
## sample estimates:
## probability of success
##          0.5691489
```

The output of this test will include the test statistic (the number of observations above 2800 gram), the p-value, and a confidence interval for the true proportion of observations that are above 2800 gram. Assuming the output p-value we reject the null hypothesis in favor of alternative hypothesis.

- c) Propose a way to compute the powers of the t-test and sign test from
- b) at some  $\mu > 2800$ , comment.

To compute the power of the t-test, we need to specify a hypothetical value of the mean birthweight, denoted by  $\mu_1$ , which is greater than 2800g. Let's say we choose  $\mu_1 = 2850$ g. Then we can use the following approach to compute the power.

For the t-test, we can compute the t-value and p-value for the null hypothesis  $H_0: \mu = 2800$ g versus the alternative hypothesis  $H_a: \mu > 2800$ g, using the formula:  $t = (\mu - \mu_0) / (sd / \sqrt{n})$ .

The p-value can then be obtained from the t-distribution with  $n - 1$  degrees of freedom. Once we have the p-value, we can compute the power of the test for the specified value of  $\mu_1$ .

For the sign test the p-value can be computed for the null hypothesis  $H_0: P(X > 2800) = 0.5$  against the alternative hypothesis  $H_a: P(X > 2800) > 0.5$  using the binomial distribution:  $p = \text{sum}(\text{dbinom}(k:n, n, 0.5))$ ,

k - number of observations greater than 2800g.

The p-value is then compared to the significance level to determine whether to reject or fail to reject the null hypothesis.

- d) Let  $p$  be the probability that birthweight of a newborn baby is less than 2600 gram. Using asymptotic normality, the expert computed the left end  $\hat{p}_l = 0.25$  of the confidence interval  $[\hat{p}_l, \hat{p}_r]$  for  $p$ . Recover the whole confidence interval and its confidence level.

If the probability that birthweight of a newborn baby is less than 2600 gram is  $p$ , then the number of newborn babies out of 188 with birthweights less than 2600 gram follows a binomial distribution with parameters  $n = 188$  and  $p$ . Assuming that the sample size is large enough, we can approximate the distribution of the sample proportion by a normal distribution with mean  $p$  and variance  $p(1-p)/n$ .

The expert computed the left end  $\hat{p}_l = 0.25$  of the confidence interval  $[\hat{p}_l, \hat{p}_r]$  for  $p$ . Since the normal distribution is symmetric, we have  $\hat{p}_r = 1 - \hat{p}_l = 1 - 0.25 = 0.75$ .

To recover the whole confidence interval, we need to find the  $z$ -value that corresponds to the left tail probability of 0.25. We can use the `qnorm()` function in R to find this value.

```
z_star <- qnorm(0.25)
z_star
```

```
## [1] -0.6744898
```

Using the normal approximation to the binomial distribution, the confidence interval for  $p$  is.

```
p_hat <- sum(dataset$birthweight < 2600) / nrow(dataset)
se_p <- sqrt(p_hat * (1 - p_hat) / nrow(dataset))
ci_p <- p_hat + z_star * se_p + c(0, 1) * 1.96 * se_p
ci_p
```

```
## [1] 0.3066602 0.3738650
```

- e) The expert also reports that there were 34 male and 28 female babies among 62 who weighted less than 2600 gram, and 61 male and 65 female babies among the remaining 126 babies. The expert claims that the mean weight is different for male and female babies. Verify this claim by an appropriate test.

To test if the mean birthweight is different for male and female babies, we can perform a two-sample  $t$ -test.

First, we can split the data into two groups based on the gender of the newborns.