ex2_cholesterol

Paul Hosek

2023-02-17

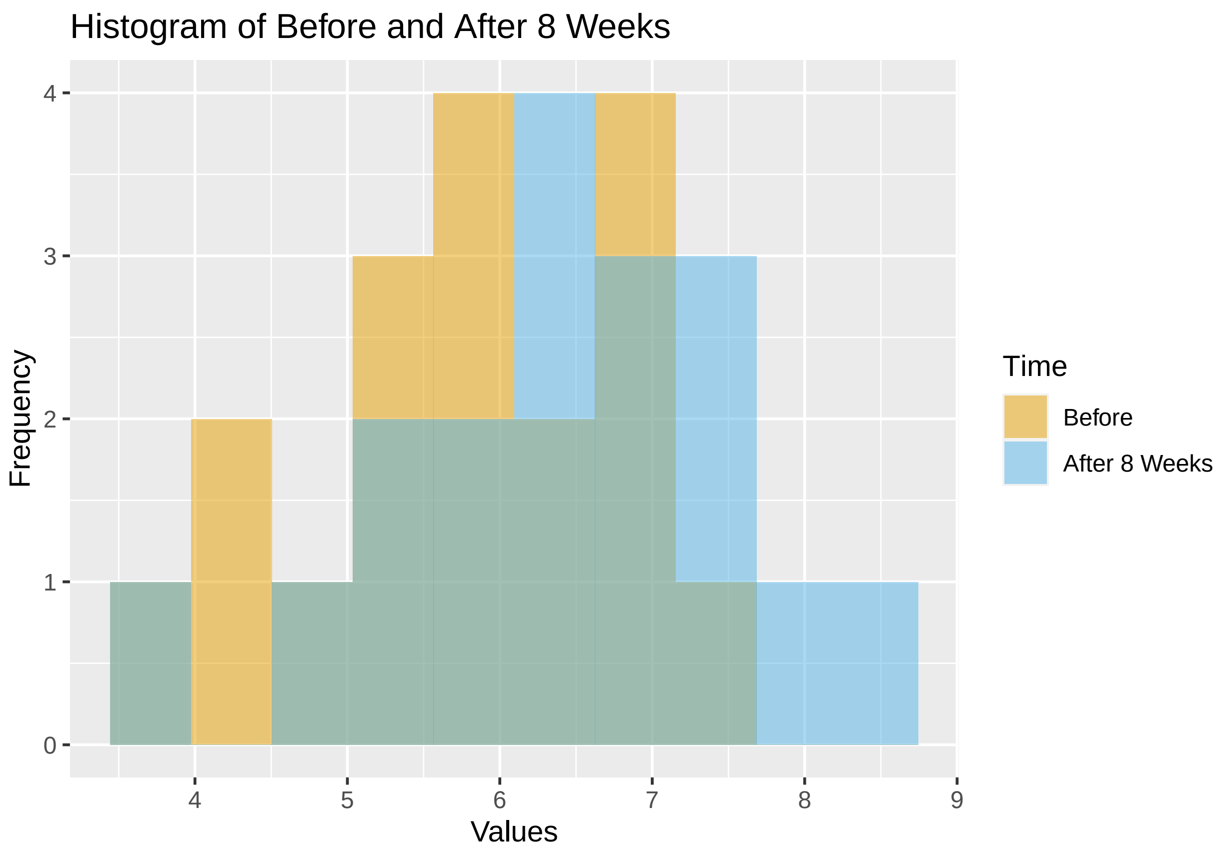
```
library(ggplot2)
library(dplyr)
library(ggpubr)
data <- read.table("data/cholesterol.txt", header = TRUE,sep = " ");
print(colnames(data))
```

```
## [1] "Before"      "After8weeks"
```

Including Plots

First plotting the discrete distibution of values, we see that values apprear approximately normally distibuted. Further, QQ-plots indicate that samples from both measurements are similarly distibuted by insepection. However, we need to confirm this with thorough statistical testing.

```
data_long <- data.frame(
  Time = rep(c("Before", "After 8 Weeks"), each = nrow(data)),
  Value = c(data$Before, data$After8weeks)
);
ggplot(data_long, aes(x = Value, fill = Time)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 10) +
  labs(title = "Histogram of Before and After 8 Weeks", x = "Values", y = "Frequency") +
  scale_fill_manual(values = c("#E69F00", "#56B4E9"), labels = c("Before", "After 8 Weeks"));
```

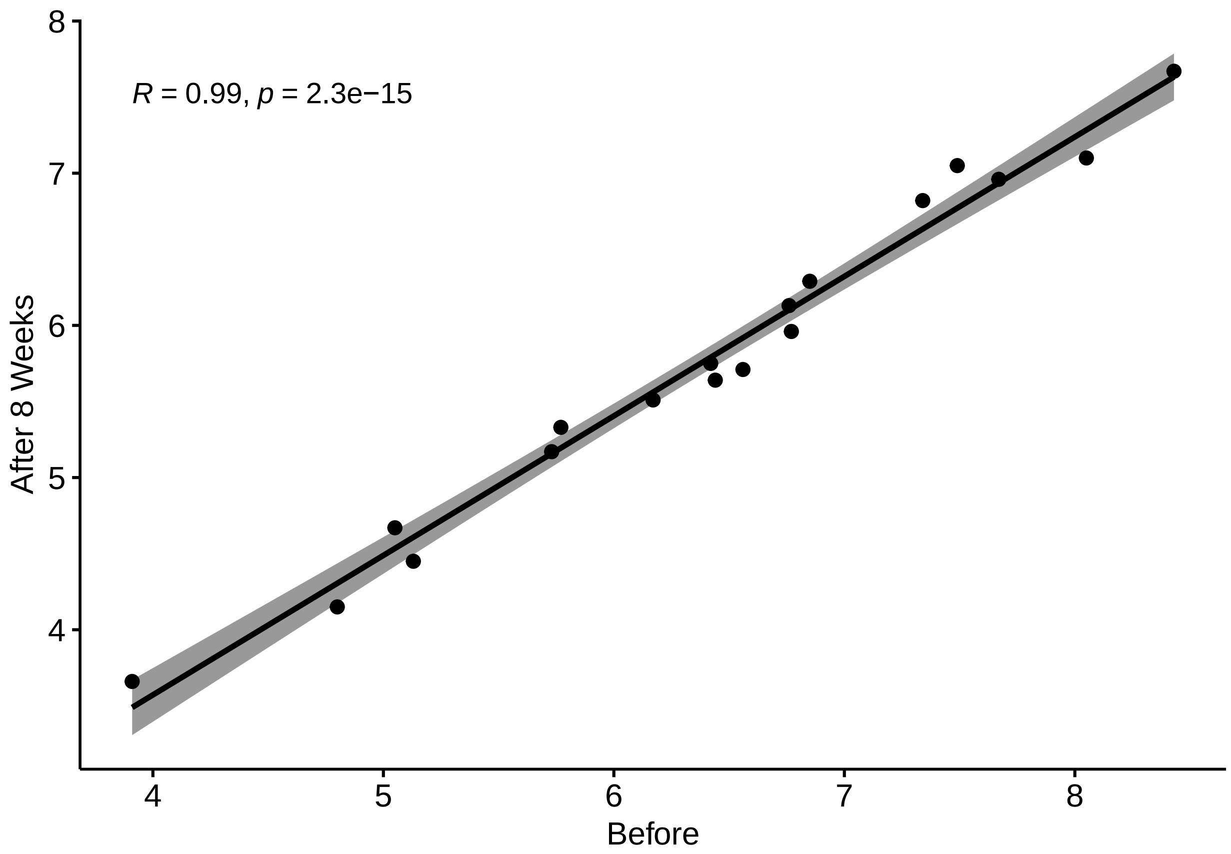


```
res1 <- shapiro.test(data$Before)$p.value
res2 <- shapiro.test(data$After8weeks)$p.value
cat("Shapiro-Wilk tests for normality indicate that the assumption of normality \n
    is met for both pre- and postmeasurement. P-values are ",res1,"for the \n
    before measurement and ", res2, "for the post measurement respectively.")
```

```
## Shapiro-Wilk tests for normality indicate that the assumption of normality
##
##     is met for both pre- and postmeasurement. P-values are  0.9674667 for the
##
##     before measurement and  0.9183031 for the post measurement respectively.
```

Further, we observe a very high correlation of the data.

```
ggscatter(data, x = colnames(data)[1], y = colnames(data)[2],
          add = "reg.line",cor.coef = TRUE, conf.int = TRUE,
          cor.method = "pearson",
          xlab = "Before", ylab = "After 8 Weeks");
```



# b) First, we use a non-parameteric, repeated measures t-test.

```
t.test(data$Before, data$After8weeks, paired = TRUE, alternative = "two.sided");
```

```
##
##  Paired t-test
##
## data:  data$Before and data$After8weeks
## t = 14.946, df = 17, p-value = 3.279e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.5401131 0.7176646
## sample estimates:
## mean difference
##       0.6288889
```

```
wilcox.test(data$Before, data$After8weeks, paired = TRUE, alternative = "two.sided");
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Before and data$After8weeks
## V = 171, p-value = 7.629e-06
## alternative hypothesis: true location shift is not equal to 0
```

The results indicate, that at $\alpha = 0.05$, there is a significant effect of the diet. However, this must not implyeffect has practical significance and is strong enough for the diet to be useful in practice.

The permutation test is applicable, since it can express any test-statistic including repeated measures and also is non-parameteric.

c)

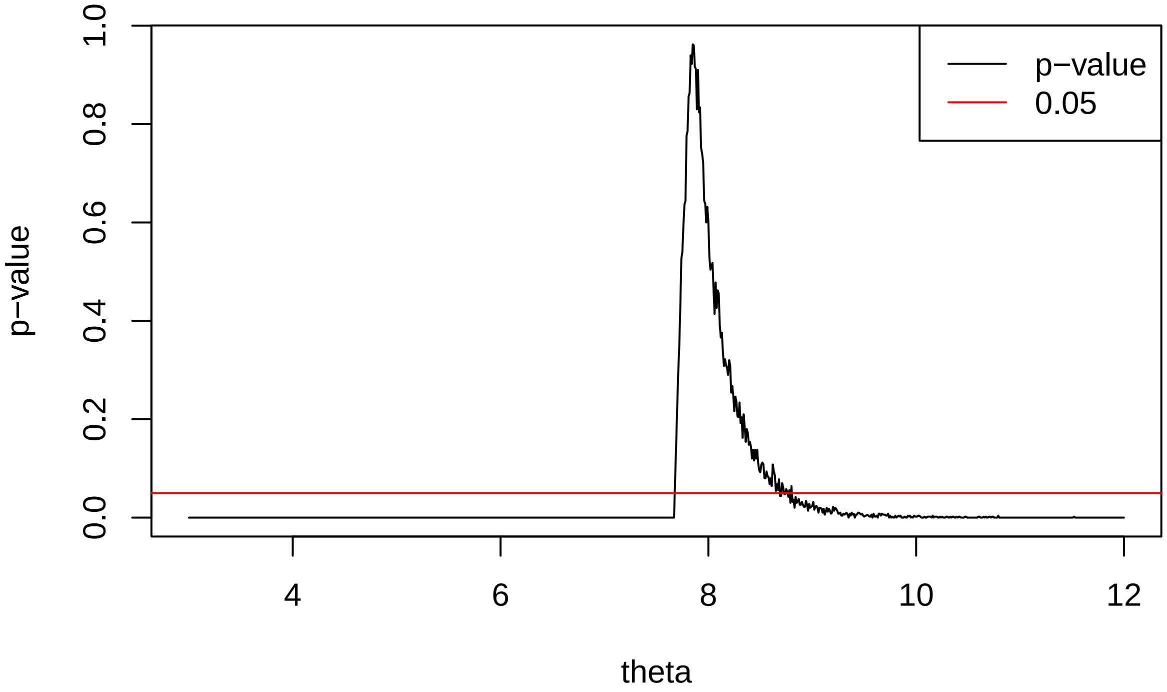
```
nsamples <- 1000;

theta <- replicate(nsamples, max(sample(data$After8weeks,size=18, replace = TRUE)));
cat("The confidence interval of the max is:",quantile(theta, c(0.025, 0.975)),"\n");
```

```
## The confidence interval of the max is: 6.96 7.67
```

This confidence interval could be improved by resampling more often. However, since the sample size is only 18, conclusions about the true population mean are limited. As such, we recommend to collect more samples.

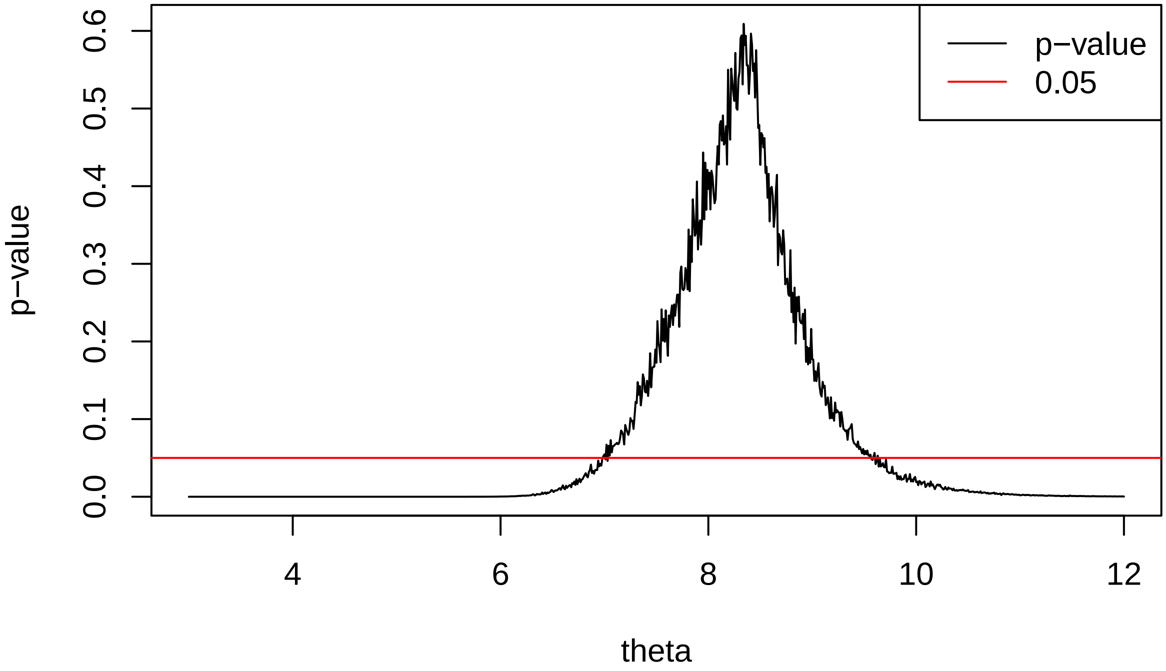
d)

```
nsamples <- 1000;
thetas <- seq(3, 12, by = .01);
t <- max(data$After8weeks)
p_vals <- numeric(length(thetas));
for (i in 1:length(thetas)) {
  res <- replicate(nsamples,max(runif(18,3, thetas[i])))
  pl=sum(res<t)/nsamples
  pr=sum(res>t)/nsamples
  p_vals[i]=2*min(pl,pr)
}

plot(thetas, p_vals, type = "l", xlab = "theta", ylab = "p-value");
abline(h = 0.05, col = "red");
legend("topright", legend = c("p-value", "0.05"),
       col = c("black", "red"), lty = c(1, 1));
```



Alternatively, we can use the Kolmogorov-Smirnov test to test if the distributions are the same. But we must be careful to not accept distributions with a maximum value smaller than the maximum of the sample, because else the maximum lies outside the given range.

```r
thetas <- seq(3, 12, by = .01);

# Loop over theta values and perform bootstrap test
ks_res <- numeric();
p_vals <- numeric(length(thetas));
for (i in 1:length(thetas)) {
  theta <- thetas[i];
  cur_unif <- runif(10000,3,thetas[i]);
  p_vals[i] <- ks.test(data$After8weeks, cur_unif)['p.value'];
}
```



e)

```
binom_e <- binom.test(sum(data$After8weeks < 6), nrow(data), alternative = "less")
cat("The p-value does not reject the H0 at alpha=0.05: p-value =",binom_e$p.value)
```

```
## The p-value does not reject the H0 at alpha=0.05: p-value = 0.8810577
```

First, we find true fraction of cholesterol levels less than 4.5.

```
prop <- mean(data$After8weeks < 4.5)
cat("The fraction of cholesterol levels <4.5 is",prop,".")
```

```
## The fraction of cholesterol levels <4.5 is 0.1666667 .
```

```
binom_e <- binom.test(sum(data$After8weeks < 4.5), nrow(data), p = 0.25, alternative = "less")
cat("No, the fraction <4.5 is not less than 25%, with p=",binom_e$p.value)
```

```
## No, the fraction <4.5 is not less than 25%, with p= 0.3056892
```