THE EFFECTS OF TWO SOPORIFIC DRUGS

In this paper, we apply a paired $t$ test to Student’s original sleep data to investigate whether two soporific drugs have different effects on patients. In the first section, we give the background and context of the data, along with some of the raw data itself. In the next section we carry out the statistical analysis, and in the last section, we discuss our conclusions.

BACKGROUND

In 1905, Cushny and Peebles [CP] conducted an experiment in which two different soporific drugs were administered to 10 patients; each patient received both drugs, but not simultaneously. The extra amount that the patient slept (beyond the patient’s baseline) was recorded. In a paper such as this, we would ordinarily describe more of how the data were collected, but because this is such a famous data set, we will not describe the process any further. Suffice it to say, this was the data analyzed by Gosset (alias “Student”) in [S], the 1908 paper in which he first developed the $t$ test. The data is part of the standard installation of R [R].

For our analysis, we look at the difference in extra sleep that each patient got with the two drugs. This gives us the following data:
<table>
<thead>
<tr>
<th>Patient</th>
<th>Difference in extra sleep (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>3</td>
<td>1.3</td>
</tr>
<tr>
<td>4</td>
<td>1.3</td>
</tr>
<tr>
<td>5</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
</tr>
<tr>
<td>7</td>
<td>1.8</td>
</tr>
<tr>
<td>8</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>4.6</td>
</tr>
<tr>
<td>10</td>
<td>1.4</td>
</tr>
</tbody>
</table>

We denote the random variable that is the difference in extra sleep (that is, extra sleep with Drug 2 minus with Drug 1) by $\Delta S$. To define this random variable more precisely, we would need to know more about how the patients were sampled; more on this at the end of the paper.

### Analysis

We use R [R] for all of the statistical computations and graphics in this paper. Our R code can be found in the accompanying script.

The non-statistical question of interest that we address is whether the two drugs administered have different effects on patients. To translate this into a statistical question of interest, we examine the random variable $\Delta S$. The statistical question of interest is whether the its random variable mean $\mu(\Delta S)$ is equal to 0 hours. If it is, then there is no difference between the soporific effects of the drugs on average; if it isn’t, then there is a difference on average.

To address the statistical question of interest, we conduct a paired $t$ significance test with null and alternative hypotheses

$$H_0 : \mu(\Delta S) = 0 \text{ hours} \quad \text{and} \quad H_{\text{alt}} : \mu(\Delta S) \neq 0 \text{ hours}.$$  

Within the context of linear models, this means that we first fit the model of class

$$\Delta S \sim 1$$
and then conduct a significance test with null and alternative hypotheses

\[ H_0 : \beta_0 = 0 \text{ hours} \quad \text{and} \quad H_{\text{alt}} : \beta_0 \neq 0 \text{ hours}, \]

where \( \beta_0 \) is the coefficient of the constant term in the fitted model.

We set the significance level at the traditional value of 0.05. Also, since we would like to know how much difference there is between the effects of the two drugs, we compute a 95% confidence interval for the random variable mean of the differences.

To conduct the significance test, we first fit the model of the class specified above in \( R \), which gives us:

\[ \hat{\mu}(\Delta S) = 1.58 \text{ hours}. \]

To check whether we can conduct statistical inference with this fitted model, we verify the sampling variability assumption of normality of the error term. We do this graphically with a normal quantile plot of the model residuals, as shown below.

Although a point or two stray from the 95% confidence bands, the pattern does not seem worrisome, especially with such a tiny sample. Assuming that the error term is normally distributed seems reasonable from this graphic.

We also verify this assumption numerically by conducting a Shapiro-Wilk test, whose null hypothesis is that the residuals are normally distributed. With this, we
find a $p$-value of 0.033, which indicates that the residuals exhibit a statistically significant departure from normality of the error term. This causes us to investigate the possible departure from normality closer in the graphic. Since the graphic doesn’t reveal anything too problematic, we continue with the analysis with the assumption that the error term is (at least approximately) normally distributed.

With the Sampling Variability Assumption verified, we conduct the desired significance test on the coefficient of the intercept in the fitted linear model, and we find a $p$-value of 0.0028. Since this is below our chosen significance level, we have found statistically significant evidence that the random variable mean of the difference in sleep amounts is nonzero. We estimate the random variable mean of the sleep amount with Drug 2 minus the sleep amount with Drug 1 to be 1.58 hours, with a 95% confidence interval (CI) from 0.70 to 2.46 hours.

**Conclusions**

Among the 10 patients tested, we found that patients using Drug 2 averaged 1.58 hours (95% CI from 0.70 to 2.46 hours) more sleep per night than patients using Drug 1. This difference is statistically significant at the 0.05 level (two-sided paired $t$ test, $p = 0.0028$).

There are naturally questions as to the extent to which these patients constitute a simple random sample, and from what population. We mention this as something that might detract from our conclusions, since the random process underlying our random variable is important in interpreting our analysis. In particular, our analysis is based on these patients giving independent observations of our random variable of sleep amounts. Also, because we don’t know exactly what population these patients are sampled from, we can’t state exactly what the random process underlying our random variable is.

In addition, we should remark again that the error term in our fitted model seemed like it might not be normally distributed, and this is an additional cause for concern as we assess the validity of our statistical inferences.

**Bibliography**

[C] John Fox. I am grateful to Douglas Bates, David Firth, Michael Friendly, Gregor Gorjanc, Spencer Graves, Richard Heiberger, Georges Monette, Henrik Nilsson, Derek Ogle, Brian Ripley, Sanford Weisberg, and Achim Zeileis

