Supplementary exercise 6.99 of IPS7e

Exploration of statistical testing, cast in a (probably fictitious) context of a rigorous training program to improve performance on a particular test (GMAT) taken by applicants to MBA programs. The intended model setup is as follows:

- a SRS (or an i.d.d. sample) of $n = 100$ students,
- assumed normal distribution $\mathrm{N}(\mu, \sigma)$ with unknown mean μ and known standard deviation $\sigma = 100$,
- interest in testing the hypotheses:
  - $\mathrm{H}_0: \ \mu = 525$ (the mean score in the general population),
  - $\mathrm{H}_a: \ \mu > 525$ (because the training program should improve scores),
- observed sample means of $\bar{X} = 541.4$ and $\bar{X} = 541.5$ for (a) and (b), respectively.

In this setup, the hypothesis is tested by the z-test,

$$z = \bar{X}/(\sigma/\sqrt{100}), \quad \text{with } P = \mathrm{P}(Z > z_{\text{obs}}).$$

The Minitab output below (from the `1-Sample Z` menu) shows the results for the situations.

Test

Null hypothesis $H_0: \mu = 525$
Alternative hypothesis $H_1: \mu > 525$

Z-Value	P-Value
1.64	0.051

Test

Null hypothesis $H_0: \mu = 525$
Alternative hypothesis $H_1: \mu > 525$

Z-Value	P-Value
1.65	0.049

As is seen, the P-value is just above 0.05 in the left display, corresponding to (a), and just below 0.05 in the right display, corresponding to (b). If the first result is labeled as "non-significant" and as indicating no proven effect of the training program, and the second result is labeled as "significant" and as indicating a proven effect of the training program, then the minuscule difference in the sample means of the two scenarios has led to a very substantial difference in the conclusions. That is clearly unreasonable, and illustrates the dangers of treating the significance level $\alpha = 0.05$ as sacred, as worded in the exercise.