

Homework One Solutions

1. The data set guinea-pigs.data gives the survival times of 72 guinea pigs after injection with tubercle bacilli in a medical experiment.

- (a) Give a 95 percent confidence interval for the mean survival time.

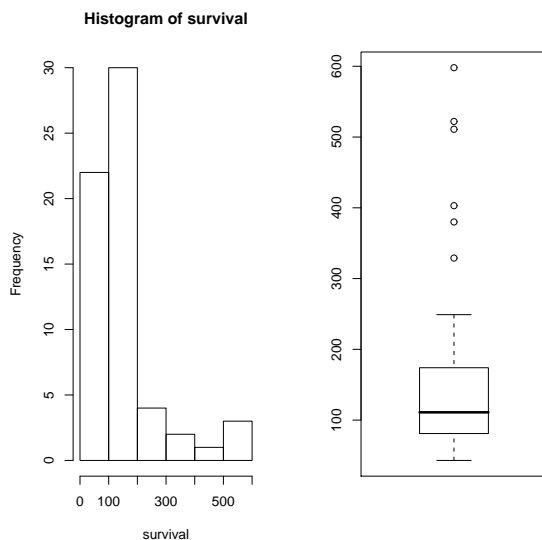
We find the 95% confidence interval for survival time to be (120.3184, 179.0687).

- (b) Check the data graphically and numerically for outliers, or anything else untoward. Report what you see. (Yes, you should ordinarily do this first, but that disrupts the flow of this problem.)

A numerical summary follows in the table:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
43.0	81.0	111.0	149.7	171.0	598.0	115.6720

We note that three guinea pigs' survival times were more than 3 standard deviations above the mean (511, 522, and 598). The mean and median are not too close together, which is possibly indicative of a problem. For a graphical summary see below:



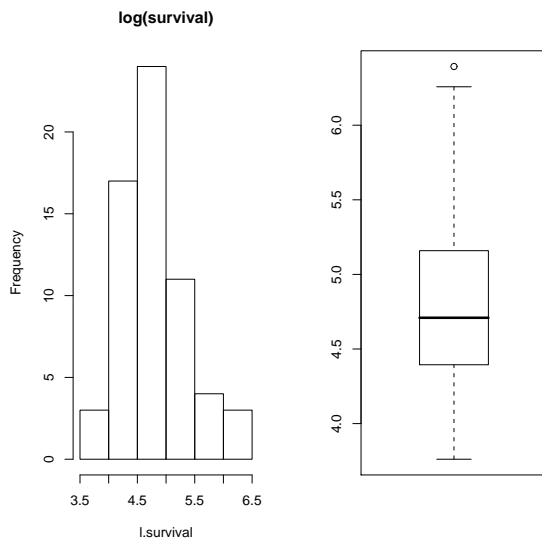
Note that the boxplot highlights several more outliers (again all stretched out to the high end), and the histogram shows the distribution of survival times is skewed to the right. It appears unlikely that this data comes from a normal population.

- (c) Take the logarithm of the data (use the log command in R) and repeat the checks above.

A numerical summary follows in the table:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
3.761	4.394	4.709	4.810	5.141	6.394	.592768

From this we see that there are no observations more than 3 standard deviations from the mean. The mean and median are now quite close. The plots for a graphical summary are below:



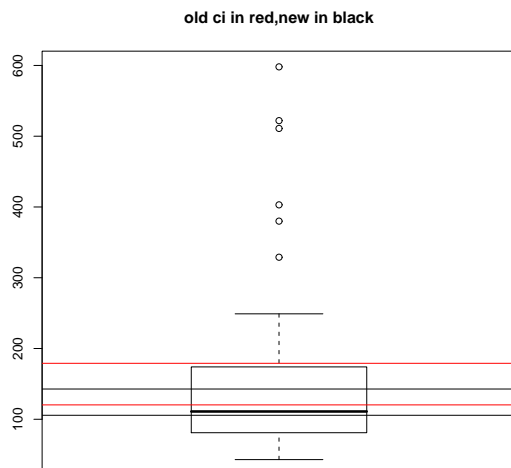
We see that overall the data seem much more symmetrical, the boxplot only highlights one outlier, and the histogram is much more like one that might have been sampled from a normal distribution.

- (d) Give a 95 percent confidence interval for the mean log survival time and transform this (use the exp command in R) back to an interval of untransformed units (i.e. undo the logarithm).

We found the 95% confidence interval for $\log(\text{survival time})$ to be (4.659913, 4.960983). The units here are $\log(\text{hours})$ or $\log(\text{original units})$, whatever the original units were, but we can convert back to original units as follows:

$$(e^{4.659913}, e^{4.960983}) = (105.6269, 142.7340)$$

The plot below shows the original data (in a boxplot) with both confidence intervals. The first one we computed is in red, and the second (untransformed from the log confidence interval) is in black. Note both that it is shorter, and at least empirically, seems more believable from looking at the pictures (the outliers haven't pulled the confidence interval up so much).



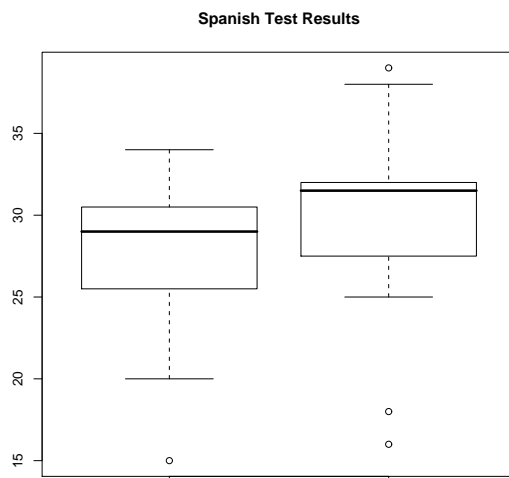
2. The dataset spanish-test.data gives pre-test and post-test scores on the MLA listening test in Spanish for 20 high school spanish teachers who attended an intensive summer course in Spanish. Does attending the institute improve listening skills?

(a) Check the data graphically and numerically for outliers or anything else untoward. Report what you see.

First a numerical summary:

pre.test		post.test	
Min.	15.00	Min.	16.00
1st Qu.	25.75	1st Qu.	27.75
Median	29.00	Median	31.50
Mean	27.30	Mean	29.75
3rd Qu.	30.25	3rd Qu.	32.00
Max.	34.00	Max.	39.00
SD	5.038	SD	5.609

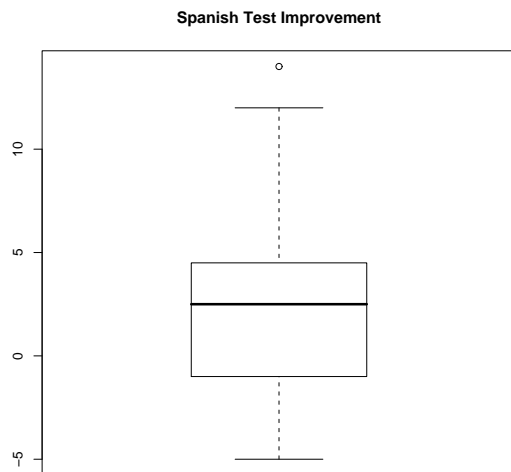
Note that for each, the mean and median are fairly close, and all observations lie within 3 standard deviations of the mean. Graphically we see that the post-test scores seem to be somewhat higher than the pre-test scores, and while there may be a few outliers, we see nothing extremely alarming.



Since this is a paired test, we'll consider the differences, and begin with a numerical summary:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
-5.00	-1.00	2.50	2.45	4.25	14.00	4.7956

none of the differences in scores lie more than three standard deviations from the mean, the mean and the median are close, and looking at the plot:



other than one possible outlier, things don't seem that alarming.

- (b) Give a 90 percent confidence interval for the mean increase in test score.

The 90 percent confidence interval for the mean increase is $(0.5958177, 4.3041823)$, so we expect scores to go up (on average) by at least $\frac{1}{2}$ point, and as much as 4.3 points.

- (c) State appropriate null and alternative hypotheses for a hypothesis test.

$H_0 : \mu_{diff} \leq 0$, the mean score does not increase

$H_1 : \mu_{diff} > 0$, the mean score does increase

Here μ_{diff} is the *post.test* minus *pre.test* mean, so the mean increase in listening test scores.

- (d) Carry out your hypothesis test and report the p-value.

We obtain a *p*-value of .017, leading us to reject H_0 in favor of H_1 and conclude that the test scores do in fact increase.

3. The data set *study-habits.data* gives results for students at a selective private college on the Survey of Study Habits and Attitudes (SSHA), a psychological test designed to measure the motivation, attitudes, and study habits of college students. Most studies have found that the mean SSHA score is lower for men than that for women.

- (a) Again check the data graphically and numerically for outliers or anything else untoward. Report what you see.

Below find numerical summaries for the male and female data:

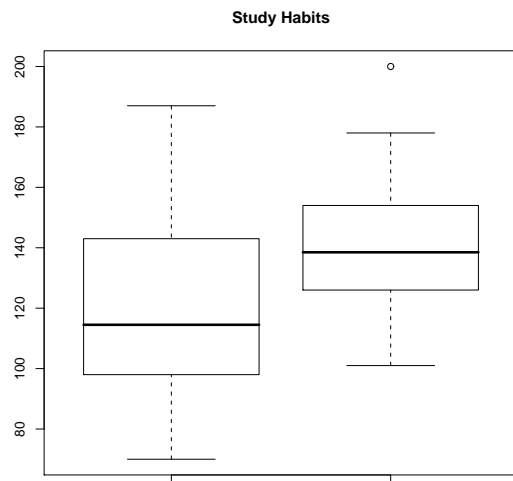
Men

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
70.0	101.0	114.5	121.2	141.5	187.0	32.85194

Women

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
101.0	126.0	138.5	141.1	154.0	200.0	26.43632

The means and medians are relatively close for each sex. All SSHA scores are within three standard deviations of the mean. See graphs below:



both sets (male and female) seem relatively evenly distributed (high and low) and we see only one potential outlier in the female data set.

- (b) Do these data support the claim that the mean SSHA score is lower for men than for women? (Use both a confidence interval and a hypothesis test to answer this question.)

$$H_0 : \mu_M \geq \mu_F, \text{ males are not lower}$$

$$H_1 : \mu_M < \mu_F, \text{ males are lower}$$

We obtain a p -value of .02358, leading us to reject the null hypothesis and conclude that female SSHA scores are indeed, on average, higher. The 95% confidence interval (one-sided) indicates that the average male score is at least 3.538 lower.