Notes and Solutions from the Book: Basic Statistics: Understanding Conventional Methods and Modern Insights by Rand R. Wilcox.

John L. Weatherwax*

Jan 12, 1997

*wax@alum.mit.edu

Chapter 2 (Basic Statistics)

Notes on the text

Notes on trimmed and Winsored statistics

In the `R` code `tmean.R` and `winsorize.R` one will find simple code to compute the trimmed mean and to return a Winsorized data set. The `R` code `chap_2_examples_N_problems.R` demonstrate their use on examples and problems from this chapter.

Problem Solutions

Problem 27

Using `tmean.R` we find 80.04762.

Problem 35/36

Using `winsorize.R` we find the Winsorized data set given by $24, 24, 25, 32, 35, 36, 36, 42, 42$. The variance of this is 51.361.

Problem 37

We would expect the Winsorized variance to be *smaller* than the sample variance. The sample variance in this case equals 81 and is indeed larger.

Problem 38

We would expect the Winsorized sample variance s_w^2 to *always* be smaller than the sample variance s^2 since when we compute the Winsorized data set we are decreasing the magnitude of the largest and smallest numbers.

Chapter 5 (Sampling Distributions)

Notes on the text

Notes on the confidence interval for the population median

In the python code `mckean_schrader.py` one will find a simple version of the McKean-Schrader estimator for the confidence interval for the population median. When this code is run from the command line it correctly duplicates Example 3 from the book.

Problem Solutions

Problem 18

See the python code `chap_5_problems.py` for a call to the python code `mckean_schrader.py` to compute the standard error of the median.

Problem 19

The previous problem has no repeated values so using the McKean-Schrader estimate of the standard error of the population median is reasonable.

Problem 20

See the python code `chap_5_problems.py` for a call to the python code `mckean_schrader.py` to compute the standard error of the median on this data set. **Warning:** The above codes produce a value of 0.5823 for the standard error where as the book claims it should be 0.97. This example has many tied values which indicates that the standard error of the median will most likely be computed poorly.

Problem 21

The previous problem has many repeated values and so we expect the McKean-Schrader estimate of the standard error to perform poorly.

Problem 22

We should probably *not* use the central limit theorem result with σ_M replaced with s_M to calculate the probability requested since the data in Problem 20 has many repeated values. In the case of repeated values it is less likely that $s_M \approx \sigma_M$.

Problem 23

Since the median is a more robust measure of the central tendency of a set of data, the fact that this given data set has a likely "outlier" with a value of 201, indicates that the given point will affect the estimate of the mean and correspondingly the estimate of the standard error of the mean. We would expect a much larger estimate of the standard error of the mean due to this point, while the standard error of the median should not be as affected. We can verify this by computing both of these estimates. In the `python` file `chap_5_problems.py` we use the code `mckean_schrader.py` to compute the McKean-Schrader estimate of the standard error of the median and the expression $s/\sqrt{n}$ to estimate the standard error of the mean. We find these two numbers given by 3.688 and 16.592 showing that indeed the standard estimate of the mean is larger. **Warning:** These numbers are slightly different than given in the back of the book.

Problem 24

Assuming normality would certainly not be a good assumption when the data values are discrete and repeated values are very likely.

Chapter 6 (Estimation)

Notes on the text

Notes on the confidence interval for the probability p of success

There were a number of computational routines developed for this chapter. In the python code `agresti_coull.py` one will find a simple version of the Agresti-Coull estimator for confidence interval for the probability of success p given the number of success X from n Bernoulli trials as discussed in the text in this chapter. In the python code `blyth.py` one will find an implementation of Blyth's method for estimating this confidence interval when the number of successes X is near 0 or n, specifically equal to one of the values 0, 1, $n-1$, or n.

Problem Solutions

Problem 1

A 0.95 confidence interval means that with probability of 95% the *true* population mean μ is contained in the interval returned.

Problem 2

When the population standard deviation σ is *known* then the random variable $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is distributed as a standard normal. Thus the value of c for which $\bar{X} \pm c\frac{\sigma}{\sqrt{n}}$ is an $1-\alpha$ % confidence interval is the one that satisfies

$$P(Z \leq c) = 1 - \frac{\alpha}{2}\,,$$

In `R` the value of c can be computed with the `qnorm` function. In the `R` script `chap_6_prob_2.R` this is demonstrated and we find the three requested values given by

```
1.281552, 1.750686, 2.326348.
```

Problem 3

We compute for $\alpha = 0.05$ the value of c for which $P(Z \leq c) = 1 - \frac{\alpha}{2}$, where $P(\cdot)$ is the c.d.f. for the standard normal. Using the `R` function `qnorm` we find $c = 1.959964$, so the 0.95

confidence interval is given by

$$\bar{X} \pm \frac{c\sigma}{\sqrt{n}} = (43.040, 46.95)\,.$$

Problem 4

This is the same as problem 3 but now $\alpha = 0.01$.

Problem 5

Given the sample mean $\bar{X}$ we will compute a 0.95 confidence interval and see if it overlaps with the proposed population mean of $\mu = 1200$. We find that with $\alpha = 0.05$ that $c = 2.575829$ and then that

$$\bar{X} \pm \frac{c\sigma}{\sqrt{n}} = (1141.833, 1158.167)\,.$$

Since this does *not* contain the value of $\mu = 1200$ there is evidence that this is not the correct value for the population mean and the above confidence interval indicates that in fact the mean is probably *less* than this value.

Problem 7 (changing the number of samples n)

Since the $1 - \alpha$ confidence interval has end points given by

$$\bar{X} \pm \frac{c\sigma}{\sqrt{n}}\,,$$

it has a width given by

$$2\frac{c\sigma}{\sqrt{n}}\,.$$

When the number of samples, n, doubles we see that this width changes to

$$2\frac{c\sigma}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\left(2\frac{c\sigma}{\sqrt{n}}\right)\,,$$

and the original interval is reduced by the factor $\frac{1}{\sqrt{2}}$. In the same way when the number of samples is multiplied by four the length of the confidence interval shrinks by a factor of two.

Problem 8

Part (a): We find $(19.9602, 19.9798)$.

Part (b): The confidence interval computed implies that there is only a 5% chance that the true population mean is outside of the above interval. This would indicate that there is evidence that the specification is *not* being met. **Warning:** This is different than in the back of the book.

Problem 11

Part (a): Since the random variable T, as specified, is given by a t-distribution with $n-1$ degrees of freedom we can calculate the requested probability using

$$P(T > c) = 1 - P(T < c) = 0.025 \Rightarrow P(T < c) = 0.975\,.$$

We can evaluate this with the `R` command $qt(0.975, 20)$ to get 2.085963.

Part (b): In this case we can directly use the `R` command `qt( 0.995, 20 ) = 2.845340`.

Part (c): To evaluate this we first use the fact that

$$P(-c \le T \le c) = P(T \le c) - P(T \le -c)\,.$$

We next note that, since $P(T \le c) = 1 - P(T \ge c)$ and that the t distribution is symmetric about the origin we can write $P(T \ge c) = P(T \le -c)$. These two observations combined show that the desired probability $P(-c \le T \le c)$ is equal to

$$\begin{aligned} P(-c \le T \le c) &= 1 - P(T \ge c) - P(T \le -c) \\ &= 1 - P(T \le -c) - P(T \le -c) = 1 - 2P(T \le -c)\,. \end{aligned}$$

Setting this expression equal to 0.9 and solving for $P(T \le -c)$ gives

$$P(T \le -c) = 0.05\,.$$

We can find the value of $-c$ where this is true with the `R` command `qt( 0.05, 20 ) = -1.724718`, thus $c = 1.724718$.

Problem 12

Part (a): We find $(19.56179, 32.43821)$.

Problem 15

We find $(10.69766, 23.50234)$.

Problem 16

We will use the version of the central limit theorem tailored for the population median. First define a random variable Z_M as

$$Z_M = \frac{M - \theta}{S_M} ,$$

where θ is the population median, M is the sample median, and S_M is the McKean-Schrader estimator of the standard error of the median. Then the central limit theorem we consider states that Z_M will be approximately normal. Thus our confidence interval for θ is given by an expression of the form

$$M \pm cS_M ,$$

where c is the $1 - \alpha/2$ quantile of the standard normal distribution. For this problem our desired probability coverage is 95% which gives $\alpha = 0.05$ and thus the value of c is given by $c = 1.959964$. Using this our confidence interval in the median is then given by

$$M \pm cS_M = 34 + 1.96 * 3 = (28.12011, 39.87989) .$$

Problem 17

We can use the python code `mckean_schrader.py` to compute the McKean-Schrader estimate of standard error of the population median. We find that it is given by 53.768. We can then use the central limit theorem for medians to derive a 99% confidence interval for the median θ. The desired confidence interval will have a form given by

$$(M - cs_M, M + cs_M) ,$$

where c is the $1 - \alpha/2$ quantile of the standard normal. When we use the data suggested for this problem we find that confidence interval give by

$$(123.5011, 400.4988) .$$

Warning: This result differs from that in the back of the book.

Problem 18

With $n = 10$ and $k = 2$ we have the confidence interval based on a binomial distribution

$$P_2 = P(2 \leq Y \leq 10 - 2) = P(2 \leq Y \leq 8) = F(8; 10, 0.5) - F(1; 10, 0.5) = 0.978515 ,$$

where $F(k; n, p)$ is the cumulative density function for the binomial density function with parameters $n = 10$ and $p = 0.5$.

Problem 19

With $n = 15$ and $k = 4$ we have we have the confidence interval based on a binomial distribution

$$P_4 = P(4 \leq Y \leq 15 - 4) = P(4 \leq Y \leq 11) = F(11; 15, 0.5) - F(3; 15, 0.5) = 0.96484\,,$$

where $F(k; n, p)$ is the cumulative density function for the binomial density function with parameters $n = 15$ and $p = 0.5$.

Problem 20

When we sort the data from Problem 14 we find that the spot of the number 88 is located at $k = 3$. Then the probability confidence for this value of k (since $n = 19$ in this case) is given by

$$P_3 = P(3 \leq Y \leq 19 - 3) = P(3 \leq Y \leq 16) = F(16; 19, 0.5) - F(2; 19, 0.5) = 0.9992\,.$$

where $F(k; n, p)$ is the cumulative density function for the binomial density function with parameters $n = 19$ and $p = 0.5$. **Warning:** While $k = 3$ is the location of the value 88 (the first element of the confidence interval) the component $X_{(n-k+1)} = X_{(19-3+1)} = X_{(17)} = 666$ which is not equal to 515 the second element of the confidence interval as it should.

Problem 21

For this experiment we see that we have $n = 15$ and $X = 5$ so we can use the Agresti-Coull method to compute a confidence interval on p the population probability of success. Using the python code `agresti_coull.py` we find one given by $(0.2365, 0.76340)$. We can also use the standard normal approximation to compute a confidence interval on the population value of p. This interval is given by

$$\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (0.0947, 0.5718)\,.$$

where c is the $1 - \alpha/2$ quantile of the standard normal.

Problem 22

For each of the given values for n and X, the central limit derived confidence interval is given by first computing

$$\hat{p} = \frac{X}{n} \quad \text{and then} \quad \text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,.$$

For example in the first case we find

$$\hat{p} = 0.2 \quad \text{and} \quad \text{SE}(\hat{p}) = 0.064\,.$$

Problem 23

From the given problem statement we find $\hat{p} = 1/10$ and a 0.95 confidence interval using the central limit theorem is given by $(-0.01759, 0.2175)$. Using the Agresti-Coull method gives $(0.053, 0.176)$ which is much smaller. **Warning:** The numbers for the central limit theorem based confidence interval are different than the ones given in the back of the book.

Problem 26

When there is only one sample, using the central limit theorem to derive a confidence interval around the population probability of success p will not result in a very good estimate of its standard error. One should use a method like Blyth's instead.

Problem 27

When we run the `blyth.py` routine we compute $(8.54\,10^{-5}, 0.00498)$ for the confidence interval. **Warning:** This is different than the result in the back of the book.

Problem 29

This is similar to the other problems we have encountered. We have $\hat{p} = \frac{5}{250}$ and we proceed to calculate the standard error around p in the usual ways.

Problem 30

In this case we cannot use a central limit based confidence interval. We must instead use Blyth's method which is tailored to cases where almost none (or almost all) of the sample are of one type. Running the python code `blyth.py` gives $(0, 0.0182)$. **Warning:** This result is different than in the back of the book.

Problem 31

An estimate of the percentage of women who have seen the advertisement is given by $\hat{p} = \frac{180}{1000}$ and a confidence interval using the central limit theorem gives $(0.1561, 0.20381)$.

Problem 32 (some affects of non-normality on the students-t distribution

One form of non-normality that is common in applications is that of skewness. When the samples of X are drawn from a distribution that is skewed then the actual distribution of T (defined via $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$) will also be skewed. As such the probability density function for these random variables T will *not* equal a symmetric student-t distribution. Because of this, the confidence interval calculated using a student-t distribution will be correspondingly incorrect.

Problem 33

When using the student-t based confidence interval one must estimate the standard deviation σ of the distribution from which samples are drawn. If outliers are relatively rare this estimated value will be smaller than the true population value, since it is unlikely that we will have observed enough samples to "see" the large variation possible from this distribution. Thus the T values we compute using this estimated standard error will have an "intrinsic" error due to the fact that the sample standard deviation will be systematically different than the population standard deviation. Basically the fact that we introduced an error in having to estimate the population standard deviation from the observed sample. The larger sample size we observe the closer our estimated standard deviation will be to the population one.

Problem 34

In the `R` code `chap_6_problems.R` we plot the given data with a boxplot. The result of running that code is presented in Figure 1. Since the 75% quantile (the upper line in the given boxplot) is at a larger distance from the median than is the 25% point (the lower line in the given boxplot). This leads us to conclude that the provided data looks to be derived from a skewed distribution. As discussed in the book, the T statistics computed from a skewed distribution will themselves be skewed. Thus using the *symmetric* student-t distribution for analysis of the confidence interval for the population mean of samples from a skewed distribution will introduce additional errors.

Problem 35

For this problem we will use the formula for the confidence interval for the 20% trimmed mean given in the book

$$\left(\bar{X}_t - c\frac{s_w}{.6\sqrt{n}}, \bar{X}_t + c\frac{s_w}{.6\sqrt{n}} \right) .$$

Where in the above expression $\bar{X}_t$ is the 20% sample trimmed mean, s_w is the Winsorized sample standard deviation, and c is the $1 - \alpha/2$ % quantile of a standard normal.

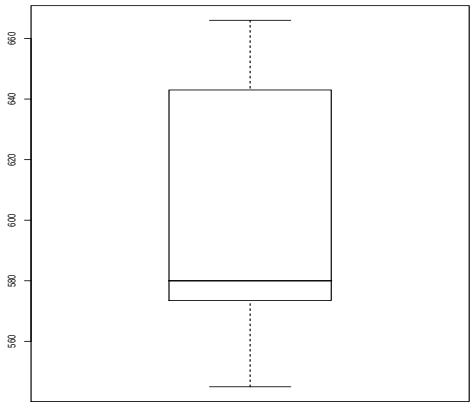

Figure 1: The box plot for the data in Problem 34.

Part (a): Using the above formula and the number specified in this part of the problem we find a confidence interval given by $(49.690, 54.309)$.

Problem 37

When we calculate the 95% confidence interval for the population mean we find $(266.865, 930.398)$ and the 95% confidence interval for the 20% trimmed mean is given by $(308.781, 580.756)$. Note that the confidence interval for the 20% trimmed mean is much smaller than that of the population mean. **Warning:** These results are different than given in the back of the book.

Problem 39/40

When we calculate the 95% confidence interval for the population mean we find $(30.734, 58.265)$ and the 95% confidence interval for the 20% trimmed mean is given by $(37.204, 52.462)$. Note that the confidence interval for the 20% trimmed mean is much smaller than that of the population mean, this is do to the potential outlier point 5. **Warning:** The result for the confidence interval for the trimmed mean are different than given in the back of the book.

Chapter 7 (Hypothesis Testing)

Notes on the text

Some important concepts/definitions introduced in this chapter are:

- Recall that α is the probability of a *Type I error*, or the error that we reject the null hypothesis H_0 when it is in fact true.
- Recall that β is the probability of a *Type II error*, or the error that we *accept* H_0 when it is false.
- The p-value is the smallest α value at which one will reject the null-hypothesis H_0.
- The *power* is the probability of rejecting the null hypothesis when it is *false*.

Problem Solutions

Sample code that computes many of the numerical values for these problems can be found in the `R` code `chap_7_problems.R`.

Problem 1

We form the statistic $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{78-80}{5/\sqrt{10}} = -1.264$. We will reject H_0 (that $\mu > 80$) when $Z \leq c$ where c is the 0.05 quantile of the standard normal. We find $c = -1.6448$. Since $Z > c$ we cannot reject H_0. We may have just observed a sample this low by chance.

Problem 2

If we want to test $H_0 : \mu = 80$ then we now need to observe that Z that with probability of 95% will be within an interval about the mean of the standard normal. This means that Z needs to be $|Z| \leq c$ where c is the $1 - 0.5\alpha$ quantile of the standard normal (otherwise we reject). This value is found to be 1.959. Since we found $Z = -1.264$ we cannot reject H_0.

Problem 3

A 95% confidence interval is given by

$$\bar{X} \pm c\frac{\sigma}{\sqrt{n}}\,,$$

where c is given as in Problem 2. This gives the interval $(74.90, 81.09)$. Since the point $\mu = 80$ is inside of this range the decision to *not* reject H_0 is consistent with this result.

Problem 4

The p-value is the smallest possible probability of a type I error given the observed value of the test statistic Z. In other words we take the critical value to be Z, then probability that another random draw from H_0 would yield a statistic Z' "beyond" this critical value is the p-value. In Problem 1 we have a one sided test, so when the critical value is taken to be $c = Z = -1.264$ as computed in that problem the p-value is given by

$$p = \text{Prob}\left\{Z' \leq c | H_0\right\} ,$$

and can be computed in `R` using the function `pnorm(-1.2649)`. When we do that we find $p = 0.1029$.

Problem 5

In this case we are looking for a two-sided test since $H_0 : \mu = 80$ and thus the p-value is given by

$$p = \text{Prob}\left\{|Z'| > c | H_0\right\} = 2 \int_{-\infty}^{c} \mathcal{N}(x; 0, 1) dx .$$

This can evaluated using the `R` command `2 * pnorm( -1.2649 )`. When we compute this we find $p = 0.205$.

Problem 6

When we compute Z in this case we get $Z = -14$ and $c = -1.6448$. Since $Z < c$ we must reject the hypothesis that $H_0 : \mu > 130$ as the evidence does not support this claim.

Problem 7

In this case all that changes is the value of the critical value c. In this case we get $c = 1.95996$ since $|Z| > c$ we reject the hypothesis that $H_0 : \mu = 130$.

Problem 8

The confidence interval for μ would be given by $(118.600, 121.40)$ thus the value of 130 is not inside that range.

Problem 9

For a *two-sided* hypothesis test, if we are not told σ than we don't know what they typical range of $\bar{X}$ is when averaging n samples from the distribution of X. If σ were a "small number" we would expect that $\bar{X}$ to be very close to μ. Or stated another way, there would be a small range about $\bar{X}$ in which the true mean μ must fall. If σ were a "large number" than we would expect a wider range around $\bar{X}$ in which μ will be found. Thus we cannot decide on whether to reject or not in a two-sided hypothesis test. For the assumption H_0 given here $H_0 : \mu < 25$ the sample value $\bar{X} = 23$ is *already* consistent with H_0 thus we cannot reject H_0 given this sample mean.

Problem 10

For the problem stated we would want to assume that $H_0 : \mu < 232$ and we would hope that our sample mean $\bar{X}$ would reject the hypothesis of H_0. We then compute Z and find $Z = 10$ and $c = 2.326$, since $Z > c$ we must reject H_0.

Problem 11

Lets take the null hypothesis to be $H_0 : \mu = 546$, then with the given numbers (and assuming a two-sided test) we find $Z = 2.124$ and $c = 1.9599$. Since $|Z| > c$ we must reject H_0 and "conclude" that $\mu \neq 546$ most likely μ is larger than that number based on the observed value of $\bar{X}$. We could compute a confidence interval around $\bar{X}$ to determine by how much (see the next problem).

Problem 12

We can determine a 95% confidence interval for μ based on the sample of $\bar{X}$, we find $(547.46, 582.53)$. Since 546 is not inside of that interval we again conclude that the true mean is larger than 546.

Problem 13

For the hypothesis test where $H_0 : \mu \geq 60$ the expression for the power is given by

$$\text{power} = 1 - \beta = P\left(Z \leq c - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) .$$

When $\alpha = 0.01$ we find $c = -2.32$ and we have $n = 25$, $\mu = 56$, $\sigma = 5$, and $\mu_0 = 60$. When we evaluate the above expression we get a power of 0.952.

Problem 14

For the hypothesis test where $H_0 : \mu \leq 100$ the expression for the power is given by

$$\text{power} = 1 - \beta = P\left(Z \geq c - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) .$$

When $\alpha = 0.025$ we find $c = 1.9599$ and we have $n = 36$, $\mu = 103$, $\sigma = 8$ and $\mu_0 = 100$. When we evaluate the above expression we get a power of 0.614.

Problem 16

Being confident in not rejecting H_0 is a case where we need to know the the power of the given test. If the power is sufficiently large we can worry less.

For the hypothesis test where $H_0 : \mu \geq 60$ the expression for the power is given by

$$\text{power} = 1 - \beta = P\left(Z \leq c - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) .$$

When we compute this we get 0.351.

Problem 18

When we change the value of n we find the power changes as

```
> ns = c(20,30,40)
> pnorm( c - ( sqrt(ns)*(mu-mu_0)/sigma ) )
[1] 0.5572501 0.7074796 0.8119132
```

Problem 19

We can increase the power by increasing the probability of a type I error α.

Problem 20

For Part (a) and (b) we cannot reject H_0 with this evidence while for Part (c) we do.

Problem 22

For Part (a) and (b) we cannot reject H_0 with this evidence while for Part (c) we do.

Problem 23

For all parts we cannot reject H_0

Problem 24

We cannot reject H_0 with the data given.

Problem 25

With $T = -2.61$, we reject H_0 with the data given.

Problem 26

With $T = 2$, we cannot reject H_0 with the data given.

Problem 27

I think there is a typo in this problem in that $\bar{X} = 76$ and $\mu = 32$, so we most certainly would reject H_0 in this case.

Problem 28

For the numbers given I get $c = 2.262$ and $T = -5.059$ which results in us rejecting H_0. These numbers are different than in the back of the book.

Problem 29

For the numbers given $c = 2.821$ and $T = 2.273$ and we cannot reject H_0.

Problem 30

When we perform hypothesis testing on the 20% trimmed mean $\bar{X}_t$ the test statistics is

$$T_t = \frac{0.6(\bar{X}_t - \mu_0)}{s_w/\sqrt{n}}\,, \tag{1}$$

where s_w is the 20% Winsorized standard deviation. The critical value c to use with this test is the $1-\alpha/2$ quantile of a Student's t-distribution with $\nu = n - h - 1$ degrees of freedom where h is the number of trimmed observations. Recall that the 20% trimmed mean is computed by first sorting the data into the sequence of observations $X_{(k)}$, computing $0.2n$, rounding this number down to the nearest integer (called g) and then we have

$$\bar{X}_t = \frac{1}{n-2g}(X_{(g+1)} + X_{(g+2)} + \cdots + X_{(n-g)})\,. \tag{2}$$

Thus the number of sampled trimmed is computed in `R` notation as `h = 2*floor(.2 * n)`.

Part (a): For the numbers given we find $h = 8$ and $c = 2.200$ and $T_t = 0.59628$, since $|T_t| \leq c$ we fail to reject.

Part (b): For the numbers given we find $h = 8$ and $c = 2.200$ and $T_t = 0.29814$, since $|T_t| \leq c$ we fail to reject.

Part (c): For the numbers given we find $h = 8$ and $c = 2.200$ and $T_t = 0.89442$, since $|T_t| \leq c$ we fail to reject.

Problem 31

The test statistics for the 20% mean is given by Equation 1, which is to be compared with a critical value from a t-distribution with $n - h - 1$ degrees of freedom, where h is given in `R` notation by `h=2*floor(.2*n)`. We first find $h = 6$, and $c = 1.833$.

Part (a): We find $T_t = 0.53333$, thus since $T_t < c$ we cannot reject H_0.

Part (b): We find $T_t = 0.2666$, thus since $T_t < c$ we cannot reject H_0.

Part (b): We find $T_t = 0.8$, thus since $T_t < c$ we cannot reject H_0.

Problem 32

We find $c = 2.570$ and $T_t = -3.103$. Since $|T| > c$ we reject H_0. This is different than the solution in the back of the book. If anyone sees anything wrong with my logic or calculations, please let me know.

Problem 33

For this problem we get $c = 2.976$ and $T_t = 0.1285$ so we fail to reject.

Problem 34

One of the main things that seem to be stressed was that the percentile bootstrap was better behaving when the underlying distribution was skewed or has heavy tails.

Chapter 8 (Correlation and Regression)

Notes on the text

Notes on least squares regression

In the python code `theil_sen.py` one will find a simple version of the Theil-Sen estimator for simple linear regression as discussed in the text on this section. This code is modeled after the code in `Rallfun-v8` from the book [1] which is an `R` version of the same robust estimator.

Notes on the population correlation ρ

From the definition of Pearson's correlation r which is the relative reduction in error in using the predictors $\hat{Y}$ as opposed to the mean $\bar{Y}$ in estimating Y given by

$$r^2 = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} ,$$

we can write this as

$$r^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} ,$$

which is the same expression as the squared multiple correlation coefficient R^2. Thus in words R^2 can be seen to be Pearson's correlation (squared) between the Y values (i.e. Y_i) and the predicted Y_i values (i.e. the $\hat{Y}_i$).

Problem Solutions

Problem 17-18

The confidence interval for β_1 is given by

$$b_1 \pm t\sqrt{\frac{s^2_{YX}}{\sum(X_i - \bar{X})^2}} ,$$

where for a $1-\alpha$ confidence interval the value of t is the $1-\alpha/2$ quantile of the t-distribution with $\nu = n-2$ degrees of freedom. For a 95% confidence interval has $\alpha = 0.05$ so we can compute the $1-\alpha/2$ quantile of the t-distribution with the `R` command

```
qt( 1 - 0.5 * alpha, n-2 )
```

See `chap_8_prob_17.R` for and `R` implemention of this problem.

Note: The solution in the back of the book for problem 18 cannot be correct since the confidence interval does not contain the estimate $b_0 = -1.5$.

Problem 19

The confidence intervals may be inaccurate for a number of reasons. One is that $n = 10$ is a very small number of points to derive accurate statistics from. Another is that if the underlying distributions are *not* normal (maybe the true distribution is highly skewed or possesses heavy tails) then the confidence intervals computed using a normal assumption may not be very reliable.

Problem 21

Part (a): We can derive an estimate of β_1 using

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{180}{60} = 3\,.$$

An estimate of β_0 can be derived using

$$b_0 = \bar{Y} - b_1\bar{X}\,,$$

from which we see that we need to compute $\bar{X}$. We can extract this later value by recalling that

$$\sum(X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2\,,$$

from which we have

$$60 = 1922 - 38\bar{X}^2 \quad \text{so} \quad \bar{X} = 7\,.$$

Thus we get that $b_0 = -1$.

Part (b): To test the hypothesis $H_0 : \beta_0 = 0$ with $\alpha = 0.02$ we compute the statistic T where T is given by

$$T = b_0\sqrt{\frac{n\sum(X_i - \bar{X})^2}{s_{Y.X}^2 \sum X_i^2}} = -0.09901\,,$$

which is to be compared to a threshold t as $|T| \geq t$, where t is the $1 - \alpha/2$ quantile of the Student's t-distribution with $\nu = n - 2$ degrees of freedom. When we want $\alpha = 0.02$ or a 98% confidence interval we find that $t = 2.43$. Thus since $|T|$ is *not* larger than t we cannot reject H_0.

Part (c): Following the box on computing confidence intervals in simple linear regression a 0.9 confidence interval for β_1 means $\alpha = 0.1$ and we would compute

$$b_1 \pm t\sqrt{\frac{s_{Y.X}^2}{\sum(X_i - \bar{X})^2}} = (0.6024587, 5.397541)\,.$$

See `chap_8_prob_21.R` for and `R` implemention of this problem.

Problem 22

We would have

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{100}{400} = \frac{1}{4} \\ b_0 &= \bar{Y} - b_1\bar{X} = 10 - \frac{1}{4}(12) = 7\,. \end{aligned}$$

The 90% confidence interval is computed with

$$b_1 \pm t\sqrt{\frac{s^2_{Y.X}}{\sum(X_i - \bar{X})^2}}\,, \tag{3}$$

where t is the $1 - \alpha/2 = 1 - 0.1/2 = 0.95$ quartile of the Student's t distribution. This number is $t = 1.6848$. Thus we get a confidence interval of $(-0.7609, 1.2609)$.

Problem 23

The 0.95 confidence interval is given by Equation 3. If $\alpha = 0.05$ the t value above is given by 2.11990 and the confidence interval is given by $(2.04, 4.15)$. Since this range is entirely greater than the value of 2 we can reject the hypothesis that $H_0 : \beta_1 < 2$ with a probability of a type I error that is less than 0.05.

Problem 24

The confidence interval for β_0 is given by

$$b_0 \pm t\sqrt{\frac{s^2_{Y.X}\sum X_i^2}{n\sum(X_i - \bar{X})^2}}\,, \tag{4}$$

where t is the $1-\alpha/2 = 0.975$ quantile of the Student's t distribution with $\nu = n-2$ degrees of freedom, this value is 2.068. Using this we compute a confidence interval of $(2.781, 9.218)$.

Problem 25

Part (a): The Pearson correlation is defined as

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2\sum(Y_i - \bar{Y})^2}}\,. \tag{5}$$

In this case these give $r = 0.8$. To test $H_0 : \rho = 0$ at the $\alpha = 0.01$ level we first compute

$$T = r\sqrt{\frac{n-2}{1-r^2}} = 6.666\,.$$

Then for $\alpha = 0.01$ we compute the $1 - \alpha/2$ quantile of the Student's t-distribution with $n - 2 = 25$ degrees of freedom. This value of t is 2.787. Since T is larger than this value we reject the hypothesis of H_0 and hence have the risk of a type I error (failure to reject H_0 at the 0.01 level). Thus for this problem we reject H_0.

Part (b): In this case $r = \frac{10}{\sqrt{16(25)}} = 0.5$ and $T = 1.0$. We compare this to $t = 3.182$. Since T is smaller we cannot reject $H_0 : \rho = 0$.

Problem 26

We compute $r = \frac{40}{\sqrt{64(100)}} = 0.5$, and then $T = r\sqrt{\frac{n-2}{1-r^2}} = 3$ and for this value of α we have $t = 1.703$. Since our T value is greater than this value of t we can reject the hypothesis $H_0 : \rho = 0$. This indicates a dependence of Y on X.

Problem 27

Part (a): Since $b_1 = \frac{r s_y}{s_x}$ is an estimate of the slope of the least squares regression line if $r > 0$ one might conclude that $b_1 > 0$ (and corresponding that $\beta_1 > 0$). As in figure 8.1 the results of potential outliers can cause r to take on any value.

Part (b): Yes, see Part (a).

Part (c): Always plot the data to verify that the least squares regression line looks reasonable.

Problem 28

Because of Equation 5 if $Y_i^* = 3Y_i$ we can show that $r^* = r$ there is no change to r under a constant multiple of Y_i.

Problem 29

Because $b_1 = r\frac{s_y}{s_x}$ if $Y_i^* = 3Y_i$ then $b_1^* = 3b_1$ i.e. the slope estimate is multiplied by 3.

Problem 30

Recall the discussion in the book on how Pearson's correlation r depends on the residual. There it was noticed that at the residual increases the correlation r decreases. This is shown in figure 8.4. Since for the second model $Y = 0.5X + 2 + 2e$, the residual will be larger (they will have a variance of 4) we expect this model to have a *smaller* Pearson correlation value of r.

Problem 31

Since we select b_0 and b_1 in $\hat{Y}_i = b_0 + b_1 X_i$ to minimize $\sum(Y_i - \hat{Y}_i)^2$ we are guaranteed that

$$\min_{(b_0,b_1)} \sum(Y_i - (b_0 + b_1 X_i))^2 \leq \sum(Y_i - \bar{Y})^2 ,$$

and the numerator in the coefficient of determination will be positive.

Problem 32

The problem with the statement just given is that it is asking to evaluate the "power" of using r to conclude that $H_0 : \rho = 0$ when we cannot reject this hypothesis. The book claims that more modern methods with greater power are available to allow one to more safely conclude that if we cannot reject $H_0 : \rho = 0$ then this hypothesis is true. One thing to note is that in the best case (i.e. no outliers) r is an linear measure of dependence. Thus if X and Y depend nonlinearly r will not capture this dependence i.e. $r = 0$ while X and Y are *not* independent.

Problem 33

Part (a): An example where a large value of r^2 does not mean that X predicts Y well is the S. Mednick study on schizophrenia where a single outlier was the cause of the strong correlation.

Part (b): Using $\hat{Y}$ is not much better than using $\bar{Y}$.

Problem 34

One might argue as follows. Because we are removing potential outliers we are removing points with large residuals. From the discussion on "interpreting r" and figure 8.4 removing points with large outliers should result in a smaller average residual and thus a larger value for Pearson's correlation r.

Problem 35

This changes the Student's t-distribution to something else.

Problem 36

The confidence interval should now depend on X in some way as in $\text{Var}(Y|X) = \sigma^2(X)$, where as in box 8.1 it is independent of X (because it assumes homoscedasticity).

Chapter 9 (Comparing Two Groups)

Problem Solutions

Problem 1-2 (Student's t-test of different means)

We implemented the two-sample Student's t-test in the python code `two_sample_student_T.py`. In that script you will find the numbers for these problems specified and the results verified.

Problem 3 (equal means?)

The T value for this problem is 9.29516, with an α value for this problem of $1.3\,10^{-12}$. We can reject the hypothesis that the means are the same.

Problem 4 (Welch's test)

We implemented Welch's test in the python code `welchs_test.py`. In that script you will find the numbers for these problems. When we run we find that using Welch's test we get a value of $W = 10.5097$ which gives a value of $\alpha = 5\,10^{-14}$.

Problem 5 (Welch's test vs. Student's t-test)

Since the sample variance between the two classes seemed so much different $s_1^2 = 4$ and $s_2^2 = 16$ we expect that Welchs' test will be the better test to use. This test gave a larger test statistics 10.5 vs. 9.25 and a smaller value for α. These indicate that it might have more power (the ability to "tell" when we should reject the hypothesis H_0).

Problem 6-7 (Student's t-test and Welch's)

In the python code `two_sample_student_T.py` and `welchs_test.py` we run tests with these numbers. For the student t-test we $T = 3.794$ with $\alpha = 0.00025$. For Welch's test we get $W = 3.7947$ with $\alpha = 0.00025828$ both of which reject the assumption that the samples have equal means.

Problem 8 (Student's t-test vs Welch's with equal variance)

When the two populations have equal variance which method one uses does not make much difference.

Problem 9-10 (using Student's t-test or Welch's)

As the sample variance are approximately equal we could use Student's t-test instead. Since we are asked to use Welch's method we will do that. I get an answer that is different than in the back of the book. I get $\nu = 29.25117$, $t = 2.0444$, and a confidence interval given by $(1.3858, 8.614141)$.

For Student t's method I again get a different result than what is found in the back of the book. I get $\nu = 30.$, $t = 2.0422724563$ and a confidence interval given by $(1.38973, 8.610261)$.

Problem 11 (training accountants)

The sample variance of these two data sets is 30159.125 and 10031.267857142857. Because they are so different we need to use Welch's method. For the data given I get $\alpha = 0.1141$ which is not sufficient to reject the hypothesis that these two samples have the same mean.

Problem 12 (hypothalamus weights)

For this problem I get that the sample mean, sample variance, and number of samples for H_0 no heart disease and H_1 to be

```
H0 (no heart disease)=  (12.063636363636363, 14.584545454545458, 11)
H1 (with heart disease)=  (15.666666666666666, 31.083809523809531, 15)
```

We see that the variance of the two classes is indeed different and Welch's test is the one to use. I however get an $\alpha = 0.031$ which would reject H_0 at the 5% level. These numerics are computed in the python code `welchs_test.py`.

Problem 13 (confidence intervals for Student's t)

From the book we have that confidence intervals using Student's t can be unsatisfactory when sampling from normal distributions with unequal variances or from non-normal distributions

with other differences like skew. Thus if the true distributions have any of these problems the confidence interval given might be incorrect.

Note: Several of the following problems are worked in the `R` script `chap_9_problems.R`.

Problem 14 (equal probabilities)

When we run that script we find $Z = 0.6182762$ with a critical value $c = 1.96$ thus we fail to reject (cannot conclude that the proportions are different).

Problem 15 (disturbing?)

When we run that script we find $Z = -0.2642733$ again with a critical value $c = 1.96$ thus we cannot conclude that the proportions are different. We get a confidence interval for the difference in proportions given by $(-0.13035626, 0.08593477)$.

Problem 16 (are these the same training program?)

We find $Z = -0.4025341$ again with a critical value $c = 1.96$ thus we cannot conclude that the proportions are different. We get a confidence interval for the difference in proportions given by $(-0.1717639, 0.1309475)$.

Problem 17 (violence in the Middle East)

For this problem I get $T = 3.3161920706374484$, $\nu = 49$, and a confidence interval for the difference in the means of $(3.2774568475731645, 13.358906788790469)$, indicating that the means do differ. The sign difference in this result vs. that from the back of the book is due to a different ordering of μ_0 and μ_1. Note that the variances of these two samples are 135.084 and 36.428 indicating that Welch's test maybe more appropriate.

Problem 18 (equal medians?)

The 95% confidence interval for this problem is $(-15.2026585, 0.2026585)$ indicating that we cannot reject the hypothesis that the medians are the same.

Problem 20 (alcohol consumption)

I *don't* get that one would reject the hypothesis that the proportions are different. I get $Z = 0.2962875$ and a confidence interval of $(-0.2441337, 0.3310902)$.

Problem 21-22 (older sisters vs. younger sisters)

I get $T_D = 5.25$ and a confidence interval of $(1.851066, 4.148934)$ arguing that there is a difference between the sisters. Problems can arise if the normality assumption is violated.

Problem 23 (improving nutrition)

I get $T_D = 2.908673$ and a confidence interval given by $(1.345246, 7.369040)$ thus we reject that they are the same mean before and after training. Unfortunately the after training mean is less than the before training mean.

Problem 24 (differing medians?)

We will use the hypothesis testing procedure for medians on the differences in test scores from the previous problem. We find a confidence interval of (1.0800728.919928) indicating that the medians *do* differ at the 95% level and thus we would reject the hypothesis that the difference scores have a median of zero.

Problem 25 (median of the differences $\neq$ difference of the medians)

The previous problem tested the median of the differences. This is *not* the same as the difference in the medians.

Problem 26 (a summary of the paired Student's t-test)

If the sample variances observed from the two sets of data was approximately the same we could apply the paired Student's t-test. If the underlying distributions are both *normal* and with equal variances then we have the optimal conditions for Student's t-test. If the distributions are not-normal but are identical Student's t-test will still perform quite well. In many cases the true Type I error probability (we reject H_0 in favor for H_1 with a probability of error given by α) will be in fact less than α. Thus if Student's t-test rejects the hypothesis $H_0 : \mu_0 = \mu_1$ then we can conclude that *something* is different between the two distributions.

Problem 27 (practical concerns with Student's t-test)

We need the two sample's variance to be equal, if they are not we may reject $H_0 : \mu_0 = \mu_1$ when in fact it is true because of this difference. If the sample variance look different (maybe differ by more than 10%) use Welch's test.

Problem 28 (using medians to determine class differences)

Method based on medians can have better power than methods based on means. This means that if in fact $H_1 : \theta_0 \neq \theta_1$ a median based method is more likely to actually reject H_0. In other words the median based methods are better at minimizing the probability of a Type II error (β) or in layman's terms

$$\beta = P\{\text{Don't reject } H_0 | H_1 \text{ is True}\} .$$

Thus if its important to be notified "immediately" when the central location has changed i.e. when H_1 is now true one should consider median based methods.

Problem 29 (the percentile bootstrap method)

Following the discussion in the book and using the numbers given in this problem we have $B = 1000$, $A = 10$, $C = 2$. Then we find

$$Q = \frac{A + 0.5C}{B} = 0.011 ,$$

$$P = \min(Q, 1 - Q) = 0.011 .$$

Then we get a p-value for testing the hypothesis $H_0 : \theta_0 = \theta_1$ given by $2P = 0.022$. Allowing us to reject H_0 at the 5% level and conclude that the medians do in fact look different.

Problem 30 (Yuen's method)

The self-awareness data is in table 9.2. In the python code `yuens_test.py` we implement Yuen's test. When we run that script we compute the value of the T_y statistics to be $T_y = -2.04424$. Changing the order of the input data will remove the negative sign in front of T_y.

References

[1] R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Statistical modeling and decision science. Acad. Press, San Diego, Calif. [u.a.], 1997.