Practice Materials for Exam 1

Up to one sample tests

Ranadeep Daw

2024-09-07

Note

I am not filling in the hypotheses and other text details. You must do it in your exams.

Output Analyzing

1. Look at the following output of a test in $\mathbb{R}$.

```
        One Sample t-test

data:  MyData
t = 0.95891, df = 251, p-value = 0.1693
alternative hypothesis: true mean is greater than 23
99 percent confidence interval:
 22.8524     Inf
sample estimates:
mean of x
 23.10238
```

A. What test am I performing here?

One sample t test. (Hint: Check the top line)

B. How many observations are there in the data?

Since in one sample t test with n observations, the degrees of freedom is (n-1). Here, n=252.

C. What is the significance (α) of the test?

$\alpha = 0.01$. Hint: Check the output that says (99 percent confidence interval).

D. Write down the null and alternate hypotheses.

$H_0: $\mu \leq 23$ vs $H_A : \mu$ > 23.

(Hint: Check the line: "alternative hypothesis: true mean is greater than 23". So, we are testing for the hypothised mean as 23 and our alternative is "$\geq$".)

E. What is the sample estimate of the parameter being tested here?

23.10 (Hint: Check the last line)

F. What is the conclusion of the test?

Since p-value (0.1693) > $\alpha = 0.01$, we can not reject H_0.

G. What is the associated confidence interval for the population parameter that is being tested?

99% CI = (22.85, ∞).

2. Now do the same for the following test. Also can you tell what the confidence interval is about?

```
        Exact binomial test

data:  122 and 234
number of successes = 122, number of trials = 234, p-value = 0.5564
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4553077 0.5868792
sample estimates:
probability of success
             0.5213675
```

Here, we are performing the sign test, i.e., H_0: median (M) = some number vs H_A: median $\neq$ some number.

(Hint: binomial test will show you number of successes, number of trials, and the line "alternative hypotheses: …. not equal to" tells you that my alternative hypothesis is two sided).

The confidence interval (0.455, 0.587) is for the probability of success p. Since 0.5 is within the interval, we can not reject H_0. Alternatively, p value < α also shows the failure to reject H_0.

Conceptual Questions

1. For a normal distribution with mean 10 and variance 25, what are the 3 quartiles and IQR?

Check the notes 03_Hypothesis_testing_updated.html, under normal distribution. The expression for first, second, and third quartile for a $\mathbb{N}(\mu, \sigma^2)$ variable are:

$Q_1 = \mu - 0.674 * \sigma$ Q_2 (median) $ = $ $Q_3 = \mu + 0.674 * \sigma$.

Here μ = 10; σ^2 = 25 => $\sigma = 5$.

So, $Q_1 = 10 - 0.674 * 5$ = 6.65 Q_2 = 10. $Q_3 = 10 + 0.674 * 5$ = 13.35

```r
10 - 0.67 * 5
```

```
## [1] 6.65
```

```r
10 + 0.67 * 5
```

```
## [1] 13.35
```

2. Suppose that for a normal distribution, the first and third quartiles are 10 and 25. What are the mean, variance, median, and IQR?

Flipping the above problem. Here you know the following:

$$Q_1 = \mu - 0.674 * \sigma = 10$$

$$Q_3 = \mu + 0.674 * \sigma = 25$$

First find the IQR. Recall, IQR = $Q_3 - Q_1$ = $25 - 10$ = 15.

Second, find the mean μ. From the above two formulas, $\mu = 1/2 * (Q_1 + Q_3) = 1/2 * (10 + 25) = 17.5$.

```r
1/2 * (10 + 25)
```

```
## [1] 17.5
```

Now, median = μ = 17.5. Lastly, from the same notes under Normal distribution, you will find the formula:

$$IQR = 1.35\sigma$$

Plug in all the values.

$$15 = 1.35 * \sigma$$

=> $\sigma = 15/1.35 = 11.11$.

```r
1/2 * (10 + 25)
```

```
## [1] 17.5
```

3. For a one-sample t-test, match the alternate hypotheses, rejection regions, and confidence intervals. Here T is the usual statistic for a one sample t-test for mean, and $t_{\alpha;\nu}$ is the usual notation for critical score from a student's t distribution.

Alternate Hypothesis	Rejection Region	Confidence Interval
1. $H_1 : \mu > \mu_0$	A. $\|T\| > t_{\alpha/2,n-1}$	I. $\left(-\infty, \bar{x} - t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}\right)$
2. $H_1 : \mu < \mu_0$	B. $T > t_{\alpha,n-1}$	II. $\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$
3. $H_1 : \mu \neq \mu_0$	C. $T < -t_{\alpha,n-1}$	III. $\left(\bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}, \infty\right)$

Hint:

i. Check the symbol of the alternate hypothesis. Rejection region should have the same sign.
ii. For one sided CI, if the alternate is $\geq$, you will see a $\inf$ in the right side of the CI. If alternate is $\leq$, you will see a $-\inf$ in the left side of the CI. For $\neq$, you will see two numbers like (c_1, c_2).

Answer:

i. 1 - B - III
ii. 2 - C - I
iii. 3 - A - II

4. What is the difference between a parametric estimation method and a nonparametric statistical method?

Parametric: We assume a distribution function and estimate the parameters within it. Example: under normality assumption, we estimate μ and σ^2.

Non-parametric: We don't assume any distribution function or parametric form of the shape of the distribution. It still has parameters to estimate.

5. Is it a population or sample?

1. **A survey of 200 households' monthly electricity usage over Pensacola.**

This looks like a sample.

2. **I want to study UWF students' average blood pressure. I collect data on all the students from UWF.**

Since I am studying only the UWF students and I have their entire data, it is a population.

3. **I want to study the average blood pressure of the Florida state university students. I collect data on all the students from UWF.**

I only have data from a subset (only UWF) from the entire population (all Florida state university students). Therefore, this is a sample.

4. **The annual median income over the counties from American Community survey data.**

Depends on the context. ACS (American Community survey) data itself is originally a sample from the total population; hence you can say it is a sample. However, in terms of using it in reanalysis projects (e.g., I want to analyze the median income over the counties of New York), this is often the largest source of data that you can find. In that context, you can treat the ACS data as a population.

Note, most variables in the ACS data come with sample estimates and a measure of sampling error.

5. **Daily temperature data over the Gulf of Mexico for 3 years.**

Again, depending on the context, there is no wrong answer. If I am interested in the ecology and environmental variables over the Gulf of Mexico and look at this 3 years of temperature data, it is a population. If we are interested in the biodiversity data over the Atlantic and only collect samples from the Gulf of Mexico, this is an example of a sample.

Data summarization

Here is a data on mean arterial pressure (MAP) by sex:

MAP	82	84	85	88	92	93	94	95	98	100	102	107	110	116	116
Sex	M	F	F	M	M	F	F	M	M	F	M	F	M	F	M

1. Calculate the mean, median, variance, standard deviation, and range of MAP.

2. Calculate the quartiles and the IQR of MAP. Draw a boxplot of MAP.

3. Calculate the upper and lower fence values for MAP. Checking by the fence values, are there evidence of outliers?

4. Calculate the mean and variance only for `Sex = M`.

Estimation and one sample test

Use the `BodyFat` data[1] data from HW1. Download it using:

```r
bodyfat_data = read.csv(url("https://hbiostat.org/data/repo/bodyfat.csv"))
```

1. Can you reasonably estimate the population mean, variance, and the first quartile of the `Density` (column name = *Density*)variable under normality assumption? Please justify your answer.

2. What are your reasonable estimates for the same under nonparametric method (or without a parametric assumption)? Please justify your answer.

3. Perform a one sample t test to test if the average biceps circumference (column name = *Biceps*) is less than 30.

4. We are interested in median biceps circumference and we want to test if it is greater than 33.8. How would you test?

---

1. https://hbiostat.org/data/↩