Chapter 1. Compute-intensive Hypothesis Testing

Shu Yang[1]

[1]*Department of Civil Engineering and Engineering Mechanics, the University of Arizona, Tucson, AZ 85719, USA*

Summary: A statistical hypothesis test is an approach of statistical inference. We can make an inference concerning the relationship between sample and population. For example, we collect two samples of data and each sample comes from a population. We would like to determine if some estimators (e.g. mean or median) in the two population are statistically equal by just checking the samples. In this chapter, we will first introduce the techniques and the concepts of classical hypothesis testing. Then, the modern technique of compute-intensive hypothesis testing will be demonstrated and performed using R language codes.

1.1 Classical hypothesis testing

1.1.1 Accuracy of mean

We usually use the terms central tendency (it is also known as location or μ) and scale (σ) to describe the shape of a given sample data with independent and identically distributed (i.i.d) random variables. We show our great interests in knowing the two parameters in a population. However, it may be practically impossible to know a population. Sampling from a population and estimating associated parameters using the sampled data offer us a reasonable way to know the population. For example, we may use various types of means (arithmetic mean, geometric mean, harmonic mean etc.) to measure the location; we may use the difference between maximum and minimum values in a sample data, variance, standard deviation, etc. to measure the scale. In most practices, the arithmetic mean and standard deviation are the popular statistical estimators to measure the population location and scale, respectively. For example, given a sample $\mathbf{x}$, N numeric elements are included in $\mathbf{x}$. Let x_i denote the ith elements in $\mathbf{x}$ and

$$\mathbf{x} = (x_1, x_2, \ldots, x_i, \ldots, x_n)$$

Equations 1 and 2 gives the sample arithmetic mean and sample standard deviation.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}} \tag{2}$$

The two equations give unbiased estimators for the population location (μ) and scale (σ).

Why we need to measure accuracy of mean, since we already know unbiased estimators of a population? As stated above, it is a perfect situation to know both location (μ) and scale (σ) of a population. However, it is practically impossible. Instead, we estimate population location and scale from samples. Due to the sampling mechanism, sample mean may not accurately represent population mean (i.e. actual mean), it is still necessary to estimate the accuracy of the estimators and show the statistical confidence for the estimated population mean. In this chapter, we only focus on measuring the accuracy of mean.

Two popular approaches to estimating the accuracy of an estimator include point estimation and interval estimation. We herein use the standard errors and margin of errors to conduct point estimation; while, confidence intervals are used to conduct interval estimations. The population standard error (se) and the

sample standard error ($\hat{se}$) are taken variations from population and sample standard deviations, respectively, and the equations are given as

$$se = \frac{\sigma}{\sqrt{n}} \tag{3}$$

$$\hat{se} = \frac{s}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{(N-1) * N}} \tag{4}$$

Since σ is unknown in most traffic studies, we simply use sample standard error instead of population standard error. This replacement is called *simple replacement rule*. In the following content, we refer standard errors to as sample standard errors, unless population standard errors are pointed out.

Significant levels are incorporated into the margin of error (me). For example, assuming a sample comes from a Gaussian distribution with N elements (N is large enough), the equation $me = z_{1-\frac{\alpha}{2}} * se$ represents the margin of error for $\bar{x}$ using the significant level α, where z represents the standardized Gaussian distribution ($N(0,1)$). In the case that N is relatively small (say, less than 30), the Student's t distribution is usually used instead of assuming the sample comes from Gaussian distributions, and then, the margin of error for $\bar{x}$ takes the form $me = t_{1-\frac{\alpha}{2},n-1} * se$. Both Student's t distribution and Gaussian distributions are symmetrically shaped. Student's t distributions have only one parameter, which is the degree of freedom. The shape of t distributions are approaching to the standard Gaussian distribution when the degree of freedom is large enough.

Confidence intervals (CI) for $\bar{x}$ with significant level α are derived from the margin of errors. Equation 5 is given to calculate confidence intervals.

$$CI = \bar{x} \pm me \tag{5}$$

More specifically, assuming that the sample $\mathbf{x}$ is independently and identically sampled from a Gaussian distribution $N(\mu, \sigma)$, the confidence interval can be determined by the sample size (N). Equations 6 and 7 shows that it uses the standard Gaussian distribution to calculate the CI when sample size is large (typically, sample size is greater than 30); t distribution is used instead when sample size is small.

with large sample size

$$CI \text{ for } \mu \in [\bar{x} - z_{1-\frac{\alpha}{2}} * se, \bar{x} + z_{1-\frac{\alpha}{2}} * se] \tag{6}$$

with small sample size

$$CI \text{ for } \mu \in [\bar{x} - t_{1-\frac{\alpha}{2},N-1} * se, \bar{x} + t_{1-\frac{\alpha}{2},N-1} * se] \tag{7}$$

Measuring accuracy of mean can help identify whether a sample mean statistically equals the corresponding population mean with certain errors. Example 1.1 demonstrates the calculations of sample mean, standard deviation, margin of error, and confidence interval.

Example 1.1

A sample with elements is sampled from $N(5, 9)$, and then calculate the mean, standard error, margin of error, and confidence interval. The sample data is listed as (1.9, -0.5, -10.2, -0.5, 5.7, -0.9, -5.3, 24.7, 8.6, 4.4, 0.1, 4.6, -5.2, 15.4, 19.4, 9.1, -6.6, -8.2, 16.4, 7.4, 6.4, -2.1, 14.2, 24, 24, 28.5, 2.9, 10.6, 14.6, 15.6, 2.3, -19, 8.2, 0.5, -0.8). The following R code shows the sampling and the calculation procedure.

```r
NumSample = 35;
# Specify a seed to ensure the following sampling procedure and
# sampled data are re-produceable
set.seed(49);
# Set the significant level to be 0.05
Alpha = 0.05;
# Draw 35 samples from N(5, 9) and round the values
SampleData = round(rnorm(NumSample, mean = 5, sd = 9), 1);
# Calculate sample mean, standard error, margin of error, and confidence interval
Sample.Mean = mean(SampleData);
Sample.SE = sd(SampleData) / sqrt(length(SampleData));
Sample.ME = qnorm(1 - Alpha / 2, mean = 0, sd = 1) * Sample.SE;
CI.Lower = Sample.Mean - Sample.ME;
CI.Upper = Sample.Mean + Sample.ME;
```

The population mean is obviously 5 in this example. The sample mean, standard error, margin of error, and confidence interval under the significant level 0.05 are 6.01, 1.84, 3.6, [2.41, 9.6], respectively. We could draw a conclusion, for example, neither the number 3.2 nor 9.4 statistically equal the population mean; while, the number 7.5, which falls into the confidence interval, statistically equals the population mean of 6.01, although a difference 1.49 between 7.5 and 6.01 can be observed.

1.1.2 One-sample t test

The one-sample t test is a statistical procedure, which is used to determine whether an observation may be equal to a population mean. The one-sample t test is one type of statistical hypothesis testing. In general, a statistical hypothesis testing consists of four primary components, including:

1. the statements of null hypothesis (H_0) and alternative hypothesis (H_a);
2. test statistic and significant level (α) selection;
3. the p-value and other measures (e.g. effect size and confidence interval) calculation; and
4. result interpretation
    - if p-value is greater than α, no evidence to reject H_0;
    - if p-value is less than α, evidence is observed to reject H_0

We use an example of traffic professionals routine work to implement the four components.

Example 1.2

A new speed limit sign (65 mph) has been placed on a freeway. Traffic professionals would like to see whether the new sign has impacts on the average driving speed. They have collected 25 spot speed data under free flow conditions. The data is listed as (66.1, 57.1, 75.4, 37.3, 76.2, 52, 43.9, 60, 74.4, 67.6, 86.7, 60.7, 47.8, 83.9, 69.8, 34.1, 67.6, 66.1, 50.9, 75.9, 34.3, 54.3, 56.3, 81.1, 46) mph.

The mean of the collected spot speed is 61 mph. We do observe a difference of 4 mph between the average speed and the posted speed limit. Is this difference statistically significant? One-sample hypothesis testing is a proper approach to answer the question. We then start to implement the hypothesis testing four components by stating the hypotheses. The null hypothesis H_0 is usually stated as *the sampled average speed ($\bar{x}$) statistically equals the posted speed limit (μ)*; while, the alternative hypothesis H_a is stated as *the average speed does not equal the posted speed limit*. The mathematical forms are written as:

$$H_0 : \mu - \bar{x} = 0 \text{ (or } \mu = \bar{x})$$

$$H_a : \mu - \bar{x} \neq 0 \text{ (or } \mu \neq \bar{x})$$

Based on the H_0, the test statistic is naturally selected as $\mu - \bar{x}$. We take the standard form of $\mu - \bar{x}$ by dividing the standard error and give the equation below. The value calculated from Equation 8 is called t-value, and the standard form of test statistic (i.e. t-value) follows the t distribution with the degree of freedom of 24. Then, p-value can be calculated by Equation 9.

$$\text{t-value} = \frac{\mu - \bar{x}}{\hat{se}} = \frac{\mu - \bar{x}}{\frac{s}{\sqrt{n}}} \tag{8}$$

$$\text{p-value} = 2 * (1 - pt(|\text{t-value}|, \text{degree of freedom})) \tag{9}$$

where, $pt(*)$ is the cumulative density function for t distribution; degree of freedom equals number of samples minus the dimension of the sample (i.e. n - 1).

Significant level (also known as the Type I error) is selected as 0.05. The calculated p-value is 0.2051. It is greater than the significant level, meaning that we do not have evidence to reject the H_0 and the average of spot speeds are statistically equal to the speed limit.

After revisiting the section *Accuracy of mean* and comparing the procedures between measuring the accuracy of mean and the one-sample t test, it is obvious that the statistical procedure of the one-sample t test is a variation of measuring the accuracy of mean. The two procedures share the same objective (i.e. test whether an observation is statistically equal to a population mean with certain errors), the same concepts (e.g., mean, standard error, significant level, etc.), and the same assumptions (e.g. samples are i.i.d variables and samples follow Gaussian distributions)

1.1.3 Two-sample t test

In addition to use the one-sample t test to determine whether average values statistically equal a specific value, determining whether two population means are equal is also important in traffic engineering, especially when conducting before-and-after analysis. For example, traffic professionals can perform a two-sample test to know if one can observe an increase in average speed after a segment of urban street is reconstructed and paved. The two-sample t test can be categorized into three tests, including the two-sample t test with equal variance, the two-sample t test with unequal variance, and the paired t test. We first introduce the two-sample t test with equal variance through demonstrating an example and implementing the four primary components in the hypothesis testing, and then the other two tests will be briefly introduced by illustrating the differences between the tests and the two-sample t test with equal variance.

Example 1.3

For the sake of traffic safety, speed limits should be varied by time of day on a segment of urban street. Traffic professionals would like to see whether the average vehicle speeds on the segment in daytime and nighttime are the same. If the averages equal, a new speed limit sign will be installed for slowing down the traffic at night. N_1 (35) sample speeds are collected during daytime; while, N_2 (20) sample speeds are collected during nighttime. The data is listed below.

Daytime: $\mathbf{x_1}$ (45.6, 50.2, 43.7, 43.2, 49.7, 45.5, 45.1, 40.6, 30.9, 43, 42.2, 47.5, 48.1, 39.5, 46.3, 41.9, 53.6, 44.6, 53.8, 44.8, 52.4, 46, 44, 46.1, 54.7, 40.5, 40.7, 55.8, 49, 35.9, 53.5, 50.8, 38.2, 42.1, 52.3) mph

Nighttime: $\mathbf{x_2}$ (44.3, 39.5, 49.3, 28.9, 49.7, 36.8, 32.5, 41, 48.7, 45.1, 55.3, 41.4, 34.5, 53.8, 46.3, 27.2, 45.1, 44.3, 36.2, 49.5) mph

We denote $\bar{x_1}$ and $\bar{x_2}$ are average speed during daytime and nighttime, respectively; and s_1 and s_2 are referred to as the speed standard deviation during daytime and nighttime, respectively. These four value are calculated

to be (45.8, 42.5; 5.7, 7.9). Once again, we do observe a difference of 3.3 mph between $\mathbf{x_1}$ and $\mathbf{x_2}$. The following equations assist in determining whether the difference is statistically significant.

Two population can be identified in this example: speed during daytime and speed during nighttime. μ_1 and μ_2 are denoted as the population means, respectively. The null hypothesis (H_0) is stated as *the speeds collected during daytime and nighttime have the same mean*; while, H_a is stated as *the two population means are unequal*. The mathematical forms are written as:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ (or } \mu_1 = \mu_2)$$
$$H_a : \mu_1 - \mu_2 \neq 0 \text{ (or } \mu_1 \neq \mu_2)$$

Similar to the procedure in the one-sample t test, after selecting $\bar{x_1} - \bar{x_2}$ as the test statistic and transform it to a standardized form, calculating the t-value, the degree of freedom, and the p-value are indispensable. Equations 10, 11, and 12 give the calculation details.

$$\text{t-value} = \frac{\bar{x_1} - \bar{x_2}}{s_p * \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{10}$$

$$s_p = \sqrt{\frac{(N_1 - 1) * s_1^2 + (N_2 - 1) * s_2^2}{N_1 + N_2 - 2}} \tag{11}$$

$$\text{p-value} = 2 * (1 - pt(|\text{t-value}|, \text{degree of freedom})) \tag{12}$$

where, $pt(*)$ is the cumulative density function for t distribution; degree of freedom $= N_1 + N_2 - 2$.

The t-value is calculated to be 1.7956; p-value is 0.0783, which is greater than the pre-selected significant level of 0.05. The p-value suggests that we do not have evidence to reject the null hypothesis. Therefore, traffic professionals have to install a new speed limit sign for slowing down the traffic during nighttime.

Similar to Example 1.2, we assume the two samples of speed data follow the Gaussian distributions (i.e. the assumption of normality). Besides, another underlying assumption is that the two samples of speed data have the same variance ($s_1 = s_2$). How about if the two groups of speed data have unequal variance ($s_1 \neq s_2$)? Or practically speaking, we have no prior reasons to believe the variances are equal. The solution is that we still keep most of the procedure and calculations, but the calculations of t-value and degree of freedom. Equations 13 and 14 give the calculation details when $s_1 \neq s_2$.

$$\text{t-value} = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{13}$$

$$\text{degree of freedom} = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{(\frac{s_1^2}{N_1})^2}{N_1} + \frac{(\frac{s_2^2}{N_2})^2}{N_2}} \tag{14}$$

Both the one-sample t test and the two-sample t test require to collect samples from populations independently. However, the paired t test is usually used when an experiment requires to collect two status from the same object. For example, vehicle maximum speed is determined by many factors (e.g. engine, tires, vehicle gross weight, etc.). A vehicle manufactory is testing a type of new engine on 20 vehicles $(x_1, x_2, \ldots, x_{20})$. They first record the max speed of the 20 vehicles, denoted as $(x_1^1, x_2^1, \ldots, x_{20}^1)$; and then they record the max speed again after replacing with the new engine, denoted as $(x_1^2, x_2^2, \ldots, x_{20}^2)$. Here, they record two groups with 20 max speed values for each. x_i^1 and x_i^2 are collected from the vehicle x_i, so the two values are *paired* (also called *matched*). Due to the difference in data collection procedure, the procedure for conducting paired t

tests are same as conducting one-sample t test, although two groups of data have to be collected. Let's denote $y = (\Delta x_1, \Delta x_2, \ldots, \Delta x_{20})$, where, $\Delta x_i = x_i^1 - x_i^2$, and simply plug y into the one-sample t test procedure to conduct the hypothesis testing.

1.1.4 One-way ANOVA

The two-sample t test is designed to determine whether two population means are equal by checking two samples. If we have K samples, we could perform $\frac{K(K-1)}{2}$ two-sample t tests to determine if there a population mean differs from others. However, the **Type I error** (α) in this case is calculated as $1-(1-\alpha)^{K(K-1)/2}$. We still would like to control the Type I at the significant level α. Therefore, the series of $\frac{K(K-1)}{2}$ two-sample t tests is inappropriate to handle K-sample tests for population mean equivalence. The one-way Analysis of Variance (ANOVA) is designed to test the hypothesis whether K population means equal.

Let's define the mathematical form of one-way ANOVA first, and then show the differences between two-sample t test and one-way ANOVA test. Assume we have K samples. kth sample has n_k elements and each element is denoted as x_k^i. For example, the third sample is $\mathbf{x_3}$, consisting of $(x_3^1, x_3^2, \ldots, x_3^{n_3})$. In total, we have $N = \sum_{k=1}^{K} n_k$ elements. $\bar{x} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}(x_k^i)}{N}$ is denoted as the mean of entire data; while, $\bar{x_k} = \frac{\sum_{i=1}^{n_k} x_k^i}{n_k}$ is the sample mean of kth sample. Table 1 shows the major differences between two-sample t test and one-way ANOVA test. Two-sample t test replies on checking the difference in sample mean of two samples. However, one-way ANOVA test is performed by checking the ratio of sample variances. Three types of variances are used in one-way ANOVA, including *between-samples variance* (SS_B), *within-samples variance* (SS_W), and *the total variance* (SS_T). The total variance (SS_T) is the summation of SS_B and SS_T. If we know two of the three variances, the rest can be easily calculated. The mathematical forms of the three variances are given in Table 1. The ratio of sample variances is called *F-ratio*, it is calculated as F-ratio $= \frac{SS_B}{K-1} / \frac{SS_W}{N-K}$.

Two assumptions are stated in two-sample t test. In addition to the two assumptions, one-way ANOVA test is associated with the third one: the assumption of homogeneous variance. It requires that the variances in K samples are statistically equal. Therefore, before performing one-way ANOVA test, we have to check the three assumptions.

1. The assumption of independence could be violated when an experiment is not well designed. If we doubt data does not follow i.i.d., it is required to look into the experiment and re-collect data.
2. The assumption of normality can be checked using the Shapiro-Wilks test or the Kolmogorov-Smirnov test. If the assumption is violated, none of one-sample, two-sample t test, and one-way ANOVA give the inaccurate p-values.
3. The assumption of homogeneous variance can be checked using F-test, the Bartlett's test, or the Levene's test. We here recommend using the Levene's test because it gives more robust results.

   Example 1.4

   Traffic professionals would like to see whether the seasonal impact on travel time can be observed. They collect four sample of travel time on a freeway segment under free-flow conditions in January, April, July, and November. Each sample has 25 travel times (in minutes). The data is list below.

   January: 17.8, 14.1, 21.5, 6.2, 21.8, 12.1, 8.9, 15.3, 21, 18.3, 26, 15.6, 10.4, 24.8, 19.2, 4.9, 18.4, 17.7, 11.6, 21.7, 5, 13, 13.8, 23.7, 9.7

   April: 23.8, 21.9, 21.8, 21.1, 18.6, 15, 18.2, 17.4, 13.9, 13.6, 18.8, 19.9, 18.2, 18.8, 16.5, 12.4, 26, 18.7, 16.3, 15.2, 18.8, 18.2, 20, 20, 17.3

   July: 17.6, 20.2, 20.7, 20.6, 7.6, 19.8, 22.2, 10.4, 32.6, 13, 10.3, 14.3, 13.6, 9, 10.5, 12.8, 13, 12.9, 20.8, 13.9, 9.9, 15.8, 20.8, 16.5, 20.3

   November: 15.6, 6.2, 10.8, 14.6, 19.5, 22.3, 20.2, 20.8, 13.4, 14.7, 14.6, 19, 14.7, 21, 4.9, 12.4, 16.7, 5.7, 17.8, 18, 10.3, 18.2, 18.2, 12.1, 21.9

Table 1: Comparison: two-sample t test vs. one-way ANOVA

	Two-sample t test	One-way ANOVA
null hypothesis	$\mu_1 - \mu_2 = 0$	$\mu_1 = \mu_2 = \cdots = \mu_k$
alternative hypothesis	$\mu_1 - \mu_2 \neq 0$	$\mu_i \neq \mu_k$, for some $i \neq k$
test statistic	difference in sample mean	ratio of sample variance
test statistic value	with equal variance: t-value $= \frac{\bar{x_1} - \bar{x_2}}{s_p * \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$ $s_p = \sqrt{\frac{(N_1-1)*s_1^2 + (N_2-1)*s_2^2}{N_1+N_2-2}}$ with unequal variance: t-value $= \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$	F-ratio $= \frac{SS_B}{K-1} / \frac{SS_W}{N-K}$ $SS_B = \sum_{k=1}^{K} n_i * (\bar{x_k} - \bar{x})^2$ $SS_T = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_k^i - \bar{x})^2$ $SS_W = SS_T - SS_B$ $\bar{x} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} (x_k^i)}{\sum_{k=1}^{K} n_k}$
null distribution	with equal variance: t-values follow t distributions with the degree of freedom of $N_1 + N_2 - 2$ with unequal variance: t-values follow t distribution with the degree of freedom of $\frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{(\frac{s_1^2}{N_1})^2}{N_1} + \frac{(\frac{s_2^2}{N_2})^2}{N_2}}$	F-ratios follow F distributions with the two degrees of freedom $K-1$ and $N-k$
p-value	two sided: $(1 - pt(\text{t-value})) * 2$	$1 - pf(\text{F-ratio})$
assumptions	assumption of independence assumption of normality	assumption of independence assumption of normality assumption of homogeneous variance

$pt(\cdot)$ is the cumulative function of t distribution
$pf(\cdot)$ is the cumulative function of F distribution

Before performing one-way ANOVA test, we use the Shapiro-Wilks test and the Levene's test to check the two assumptions. It shows the four samples follow Gaussian distributions and statistically have the same variance. We denote μ_1, μ_2, mu_3, and μ_4 to the population mean of travel time in the four months, respectively. The null hypothesis (H_0) is stated as $\mu_1 = \mu_2 = \mu_3 = \mu_4$; while, the alternative hypothesis (H_a) is stated as $\mu_i \neq \mu_j$, where $i, j \in [1, 4]$ and $i \neq j$. We then follow the basic hypothesis testing procedure to sequentially calculate SS_B, SS_W, SS_T, F-ratio, and p-value. The final p-value is 0.138, which is greater than the significant levels of both 0.05 and 0.1. We draw a conclusion that we do not have evidence to reject the null hypothesis under both significant levels, indicating that seasonal impact is not significant in travel time.

Note that one-way ANOVA is generalized version of two-sample t test. We can also apply the one-way ANOVA on two samples. This implies that the selection of test statistics is flexible and we could use either the difference in sample mean or the ratio of variances to determine whether two population means are equal.

1.1.5 Summary of classical hypothesis testing

Starting from the difference in concept between population and sample, the importance and key calculations of measuring the accuracy of mean was introduced. Standard errors, margin of erros and confidence interval for the population mean (μ) help measure the accuracy. The one-sample t test, which is commonly used to

determine whether a value comes from a population with a specific mean with certain errors, is considered as a variation of *measuring accuracy of mean*. Four primary components are summarized to conduct the one-sample t test. The two-sample t test sequentially executes the same components. Equations of calculating t-values and p-values for different t tests were given. The following sections give more details concerning other important concepts in these t tests.

1.1.5.1 Beyond two-sided t testing

The alternative hypotheses (H_a) in both Examples 1.2 and 1.3 are inequation-based. We call this type of statistical test *two-sided testing*. However, we can modify (H_a) according to specific practice requirements. In Example 1.3, we could state H_a as $x_1 - x_2 < 0$ (or $x_1 < x_2$), or $x_1 - x_2 > 0$ (or $x_1 > x_2$). This is called *less-than t test* and *greater-than t test*, respectively. The procedures and calculations for the two tests are kept the same but the calculations of p-values. Equations 15 and 16 are used to calculate the corresponding p-values in t tests.

less-than t test

$$\text{p-value} = pt(\text{t-value}, \text{degree of freedom})) \tag{15}$$

greater-than t test

$$\text{p-value} = 1 - pt(\text{t-value}, \text{degree of freedom})) \tag{16}$$

For example, a *one-sample less-than t test* can be conducted by calculating Equations 8 and 15; while, Equations 10, 11, and 16 can be used to conduct a *two-sample greater-than t test with equal variance*.

1.1.5.2 Moving ahead to compute-intensive hypothesis testing

Several assumptions have to be met when conducting t tests. One of the assumptions in t tests is the normality assumption. Data or samples should have a Gaussian distribution. However, in most traffic studies, we have no prior reasons to believe a traffic measure follows a Gaussian distribution. For example, previous studies have showed that travel times have a multimode shape. In addition, the selection of test statistics (e.g. the mean difference between two groups) are limited. In abovementioned t tests, the differences in means are the only choice, and then the selected test statistic or standardized test statistic has to follow either the standard Gaussian distribution or t distributions. However, the difference in means may not be an appropriate choice due to specific practice purposes. We could more flexibly use median or trimmed-mean as test statistics, and thus, these test statistics may not follow t distributions any more. Compute-intensive hypothesis testings provide a general hypothesis testing framework with a flexible choice for test statistics and weaken the normality assumption.

1.2 Compute-intensive one-sample test

We introduced the basic concepts of statistical hypothesis testing. Estimating standard errors of an estimator serves as the important and fundamental step in the testing. The classical one-sample test replies on the estimation on confidence interval with a given significant level (α) of population mean (μ). However, we do assume sample data comes from a Gaussian distribution and use the difference between the sample mean and population mean as the test statistic. However, we may not have prior evidences to meet the normality assumption with limited selections of test statistics. The bootstrap technique, which utilizes the power of computing, is able to give the accuracy of any estimators or test statistics.

1.2.1 Introduction to bootstrap

Before we jump into the bootstrap (Efron and Tibshirani (1994)), three concepts concerning sampling are required to be clarified. *Sampling* is defined as the statistical procedure of drawing sample data from a population. *Resampling* is defined as the procedure of drawing sample data from a given observed sample. Resampling can be categorized as *resampling with replacemet* and *resampling without replacement*. The bootstrap uses the resampling with replacement. For example, given 30 observed travel times on a freeway segment, denoted as $\mathbf{t} = (t_1, t_2, \ldots, t_{30})$, a possibility of resampling with replacement of t could be $\mathbf{t^{*1}} = (t_5, t_{30}, t_5, t_{27}, t_8, t_{16}, t_8, \ldots)$. Note that the number of travel times in $\mathbf{t}$ and $\mathbf{t^{*1}}$ should equal.

Mathematically, if a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is observed, and after $\mathbf{B}$ times of resampling with replacement from $\mathbf{x}$, we obtain $(\mathbf{x^{*1}}, \mathbf{x^{*2}}, \ldots, \mathbf{x^{*B}})$, where, $\mathbf{x^{*b}} = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$ is called as a *bootstrap sample* and the size of $\mathbf{x^{*b}}$ is the same as the original data $\mathbf{x}$. Now we can select a statistic of interest denoted as SI, and then apply the statistic of interest on $\mathbf{x^{*b}}$. We obtain a series of data $SI(\mathbf{x^{*1}}), SI(\mathbf{x^{*2}}), \ldots, SI(\mathbf{x^{*B}})$. The series of data is called *bootstrap replication*. We are able to estimate the accuracy of SI using bootstrap replication by evaluating Equations 17 and 18.

$$\hat{se_{boot}} = \sqrt{\frac{\sum_{b=1}^{B}(SI(\mathbf{x^{*b}}) - \bar{SI})^2}{B-1}} \tag{17}$$

$$\bar{SI} = \frac{\sum_{b=1}^{B} SI(x^{*b})}{B} \tag{18}$$

Example 2.1

300 sample data are generated from *N(5, 9)*. The population mean, median, and standard deviation are obviously 5, 5, and 9, respectively. We then calculate the observed sample mean, median, and standard deviation, and also estimate the standard errors of the three estimators. The corresponding R code is given below.

```
# Estimating standard error given a sample data and a statistic of interest
Boot.Se = function(observedData, numBoot, si)
{
    ObservationSize = length(observedData);
    BootReplication = c();

    for(i in 1: numBoot)
    {
        set.seed(i);
        BootSample = sample(observedData, size = ObservationSize, replace = T);
        BootReplication[i] = si(BootSample);
    }

    Mean.SI = mean(BootReplication);
    BootSE = sqrt(sum((BootReplication - Mean.SI)^2) / (numBoot - 1));
    return(BootSE);
}

# Test the function Boot.Se()
# Draw 300 samples from N(5, 9) and round the values
NumSample = 300;
ObservedData = round(rnorm(NumSample, mean = 5, sd = 9), 1);
NumBoot = 5000;
```

```
BootSE.Mean = Boot.Se(ObservedData, NumBoot, si = mean);
BootSE.Median = Boot.Se(ObservedData, NumBoot, si = median);
BootSE.Sd = Boot.Se(ObservedData, NumBoot, si = sd);
```

Since specific random seeds are set in the R code, we determinedly get the standard errors of the three estimators (i.e. sample mean, sample median, and sample standard deviation) 0.54, 0.72, and 0.43. For another example, we also repeatably use the R function *Boot.Se()* to determine the impact of sample size on the accuracy of the three estimators. Table 2 shows the results. It is not surprisingly that we observe a trend of decreases in the standard errors when the sample size increases. However, the increase in sample size may not guarantee an improvement of accuracy of estimators. For example, we get the the standard error of 0.59 using 50 sample data and standard deviation as the statistic of interest; meanwhile, the standard error is slightly increased when 100 sample data is plugged in. However, both of the standard errors remain at a low level. Note that the results may vary on your computer but the results keep the same.

Table 2: Sample size vs. accuracy of estimators

sample size	10	50	100	1000	∞
mean	2.18	1.03	0.94	0.29	0.00
median	3.48	1.3	1.25	0.38	0.00
standard deviation	1.14	0.59	0.75	0.21	0.00

1.2.2 Bootstrap confidence intervals

After standard errors of a statistic of interest is estimated, we switch to estimate confidence intervals. Three types of confidence interval are introduced, including percentile intervals, standard normal intervals, and improved intervals called bias-corrected and accelerated (BCa) confidence intervals (Efron and Tibshirani (1994)). Given a raw data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and a selected statistic of interest (SI), we are able to obtain the bootstrap replication $SI(\mathbf{x}^{*\mathbf{1}}), SI(\mathbf{x}^{*\mathbf{2}}), \ldots, SI(\mathbf{x}^{*\mathbf{B}})$, where, SI is the statistic of interest; $\mathbf{x}^{*\mathbf{b}}$ is the bootstrap sample. Let's denote $SI(\mathbf{x}^{*\mathbf{b}})$ as $\theta(\hat{b})$. The bootstrap replication can be written as $\hat{\theta} = (\theta(\hat{1}), \theta(\hat{2}), \ldots, \theta(\hat{B}))$. A empirical probability distribution can be generated from $\hat{\theta}$. The percentile interval is produced by jointly using the empirical distribution and the significant level (α). Equation 19 is given to calculate the percentile interval of $\hat{\theta}$.

$$[\hat{\theta}_{lower}, \hat{\theta}_{upper}] = [\hat{\theta}^{(\alpha/2)}, \hat{\theta}^{(1-\alpha/2)}] \tag{19}$$

where, $\hat{\theta_{lower}}, \hat{\theta_{upper}}$ mean the lower and upper bound of the intervals; $\hat{\theta}^{(\alpha)}$ is the 100αth percentile of $\mathbf{B}$ bootstrap replication ($\hat{\theta}$); α is the significant level.

If we assume that the bootstrap distribution follows a Gaussian distribution, the confident interval regarding $\bar{\hat{\theta}}$ can be estimated by a standard normal distribution (z). Equations 20 and 21 give the calculation of the normal interval.

$$[\hat{\theta}_{lower}, \hat{\theta}_{upper}] = [\bar{\hat{\theta}} - \hat{se_{boot}} * z_{1-\frac{\alpha}{2}}, \bar{\hat{\theta}} + \hat{se_{boot}} * z_{1-\frac{\alpha}{2}}] \tag{20}$$

$$\bar{\hat{\theta}} = \sum_{b=1}^{B} \theta(\hat{b}) \tag{21}$$

where, $\hat{se_{boot}}$ is given by Equation 17. Equation 21 mathematically equals Equation 18, meaning the average of bootstrap replications.

Bias-corrected and accelerated (BCa) confidence intervals have been proposed to improve the percentile intervals. The improved percentile intervals can be calculated using Equations 22, 23, 24, 25, and 26.

$$[\hat{\theta}_{lower}, \hat{\theta}_{upper}] = [\hat{\theta}^{(\alpha_1)}, \hat{\theta}^{(\alpha_2)}] \tag{22}$$

$$\alpha_1 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha/2)})}) \tag{23}$$

$$\alpha_2 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha/2)})}) \tag{24}$$

$$\hat{z}_0 = \Phi^{-1}(\frac{\#\{\hat{\theta}(b) < \bar{\hat{\theta}}\}}{B}) \tag{25}$$

$$\hat{a} = \frac{\sum_{b=1}^{B} (\bar{\hat{\theta}} - \hat{\theta(b)})^3}{6(\sum_{b=1}^{B} (\bar{\hat{\theta}} - \hat{\theta(b)})^2)^{1.5}} \tag{26}$$

where, $\hat{\theta}^{(\alpha)}$ is the 100 αth percentile of the bootstrap replication; $z^{(\alpha)}$ is the 100 αth percentile of the standard Gaussian distribution; and $\Phi(\cdot)$ is the standard Gaussian cumulative density function.

We continue to utilize the 300 sample data generated from *N(5,9)* in Example 2.1, and plug the data into equations of calculating the three confidence intervals with the three statistic of interest (i.e. mean, median, and standard deviation). Figure 1 shows the confidence intervals of the three statistical estimators. Two significant level (0.05 and 0.1) are applied. Not surprisingly, the ranges of confidence intervals with $\alpha = 0.05$ are greater than those of confidence intervals with $\alpha = 0.1$. Recall the sections *Accuracy of mean* and *One-sample t test*, estimating confidence intervals are the critical step prior to determining whether a value equals a population with a specific mean. Next section will give the details of conducting the one-sample test using the bootstrap.

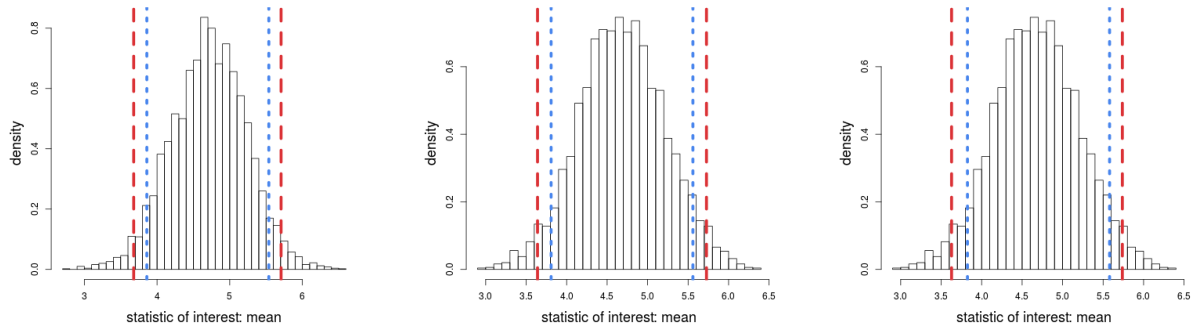
1.2.3 One-sample test using bootstrap

Three types of confidence intervals were introduced in the previous section. According to the nature of the one-sample test, we can conduct the one-sample test using the bootstrap by just checking if a value falls into the confidence intervals.

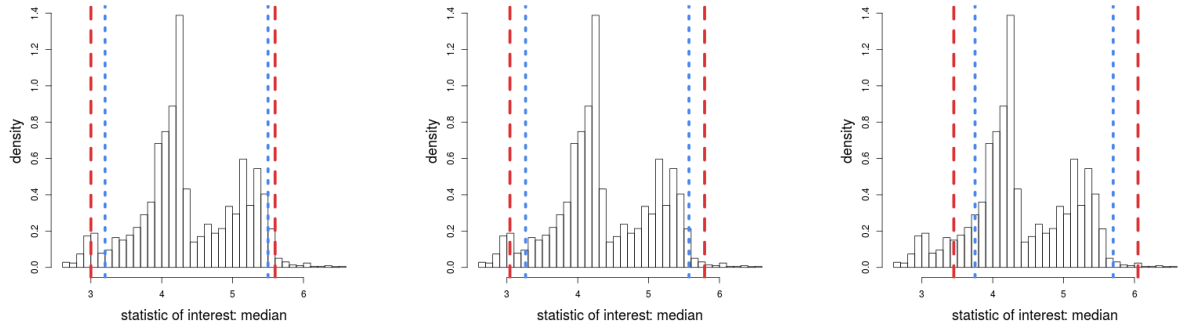
Example 2.2

Revisit Example 1.2. We continue to use the data in Example 1.2 to determine whether the average of collected speed data is statistically equal to the posted speed limit (65 mph). Both the siginificant levels of 0.95 and 0.9 are used to estimate the confidence intervals. The improved interval is used in this example. A nature selection of the test statistic would be the average of speed. In addition to the mean, we also select the median of speed as the test statistic. Therefore, the null and alternative hypotheses of $H_0 : \bar{x} - \mu = 0; H_1 : \bar{x} - \mu \neq 0$ and $H_0 : median(x_1) - \mu = 0; H_1 : median(x_1) - \mu \neq 0$ are setup for the corresponding test statistics. Again, let's denote the collected data as $\mathbf{x}$ and the speed limit as μ.

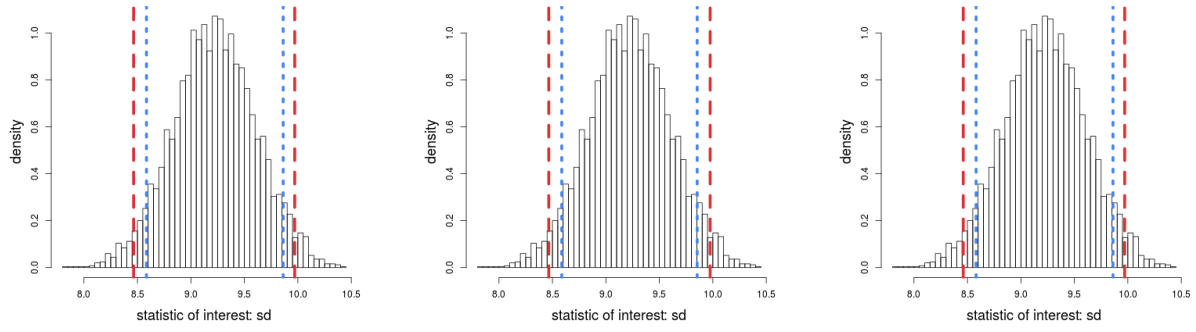
Since the test statistics are mean($\mathbf{x}$) - μ and median($\mathbf{x}$) - μ, we are required to estimate the distributions under the null hypothesis using bootstrap technique. Figure 2 shows the null hypothesis distributions and the confidence intervals associated with both $\alpha = 0.05$ and $\alpha = 0.1$. The difference under the null hypothesis (i.e. 0) falls into both intervals, indicating that we do not have evidences to reject H_0 and both average and median of the collected speed statistically equal 65 mph. This conclusion corresponds with the conclusion we drew in Example 1.2.



(a) percentile interval (SI: mean) (b) normal interval (SI: mean) (c) improved interval (SI: mean)

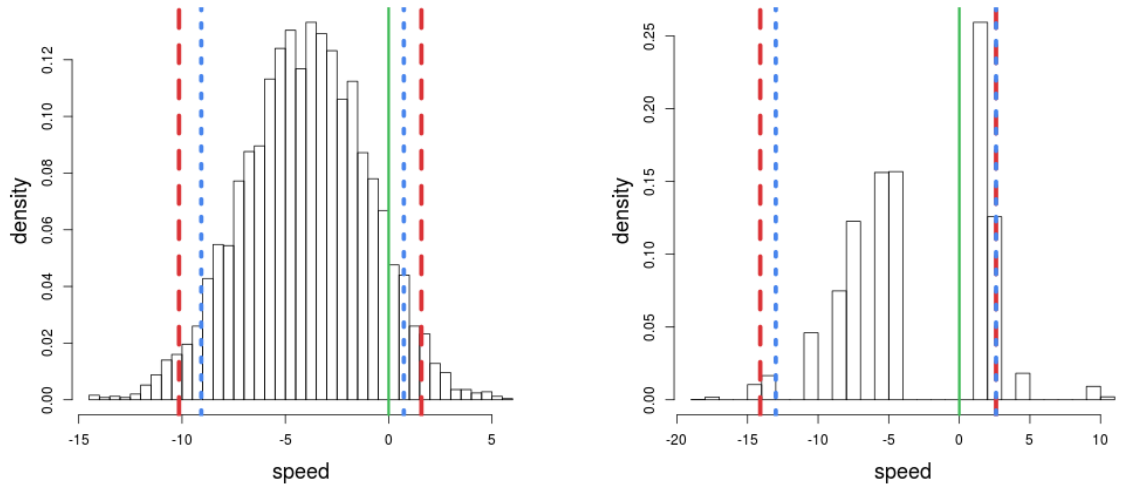


(d) percentile interval (SI: median) (e) normal interval (SI: median) (f) improved interval (SI: median)



(g) percentile interval (SI: sd) (h) normal interval (SI: sd) (i) improved interval (SI: sd)

Figure 1: Three types of confidence intervals for the three estimators. Dash lines are the confidence bounds when $\alpha = 0.05$; Dot lines are the confidence bounds when $\alpha = 0.1$



(a) test statistic: mean

(b) test statistic: median

Figure 2: Revisit Example 1.2: test statistic empirical distributions. Dash lines are the confidence bounds when $\alpha = 0.05$; Dot lines are the confidence bounds when $\alpha = 0.1$; Solid lines are the posted speed limit

In addition to the intuitive approach to determine the statistical significance of observations, we can also calculate p-values to conduct one-sample test using bootstrap. Equations 27 and 28 are given below to calculate p-values.

$$\text{p-value} = 2 * (1 - G(0)) \tag{27}$$

Since $G(0)$ can be approximately estimated as $\frac{\#(\hat{\theta}<0)}{\#(\hat{\theta})}$, so Equation 27 can be written as:

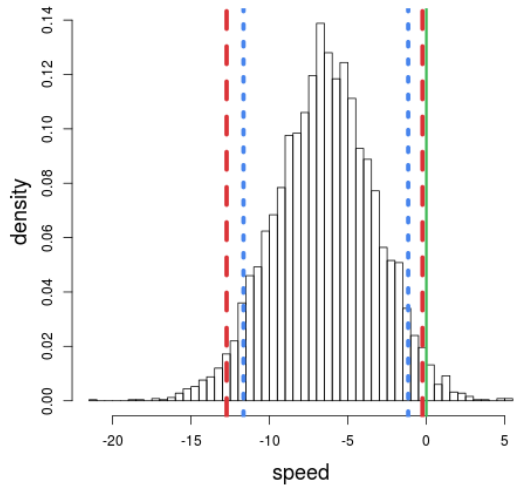
$$\text{p-value} = 2 * (1 - \frac{\#(\hat{\theta} < 0)}{\#(\hat{\theta})}) \tag{28}$$

where, $\hat{\theta}$ is bootstrap replication calculated by a specified test statistic; $G(\cdot)$ is the empirical cumulative distribution function of $\hat{\theta}$; $\#(\cdot)$ is to calculate the size of given conditions.

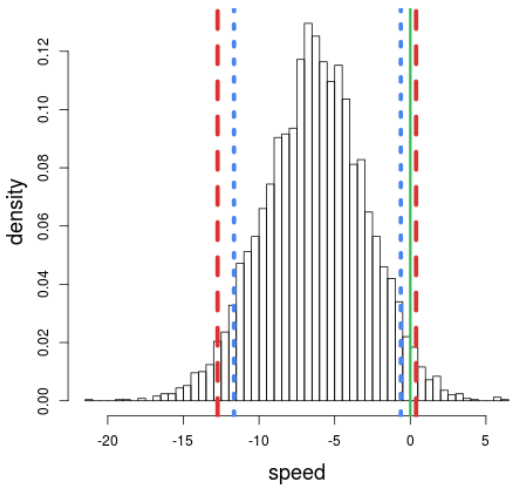
Example 2.3

We collect 25 sample speed in Example 1.2. Traffic professionals count another two vehicle speeds (25.3 and 32.6 mph). Thus, they have 27 samples in total. Obviously, the additional two speeds are far away from the speed limit (65 mph). The reason for the low speeds remains unknown, but we have no reasons to treat the two speeds as outliers. Now we use the 27 samples to conduct both one-sample t test and one-sample bootstrap test to determine if the vehicles follow the speed limit. Due to the relatively small values, besides mean($\mathbf{x}$) - μ used in Example 2.2, we also select 5% trimmed mean($\mathbf{x}$) - μ, 10% trimmed mean($\mathbf{x}$) - μ, and median($\mathbf{x}$) - μ as the test statistics in order to mathematically reduce the impact of the small values.

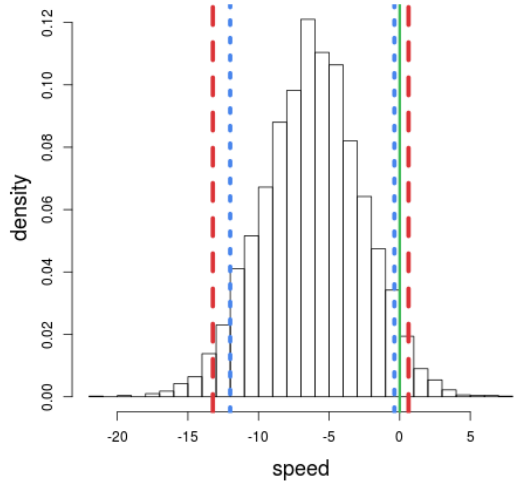
First, we give result of the one-sample t test: 0.063. The value is greater than $\alpha = 0.05$. However, the one-sample bootstrap test gives the p-values 0.038, 0.056, 0.075, and 0.535. Figure 3 and Table 3 demonstrate the detailed results. Under the same test statistics of mean($\mathbf{x}$) - μ, one-sample t test showed the average speed had no statistical differences with the speed limit; while, one-sample bootstrap test showed there was a difference and then we had the evidence to reject H_0. The result under 5% trimmed mean($\mathbf{x}$) - μ was at



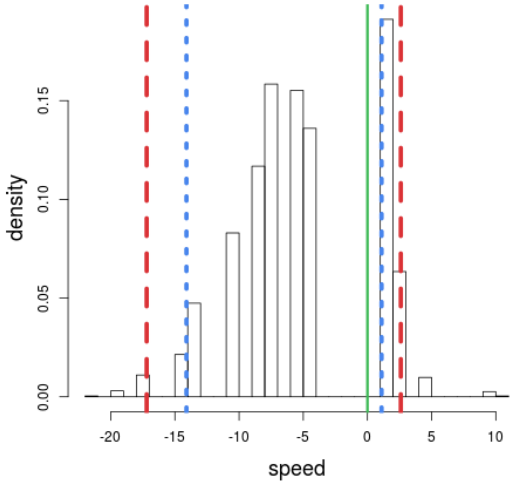
(a) test statistic: mean - μ



(b) test statistic: 5% trimmed mean - μ



(c) test statistic: 10% trimmed mean - μ



(d) test statistic: median - μ

Figure 3: One-sample test using bootstrap when extreme values are collected. Dash lines are the confidence bounds when $\alpha = 0.05$; Dot lines are the confidence bounds when $\alpha = 0.1$; Solid lines are the speed limit

Table 3: p-values calculation: bootstrap vs. t test

test statistics in one-sample bootstrap test			
mean	5% trimmed mean	10% trimmed mean	median
0.038	0.056	0.075	0.535
one-sample t test (under test statistic of mean and normality assumption)			
0.063			

borderline, we need either to draw conclusions with concerns or collect more data to start over the statistical procedure. When we switch to 10% trimmed mean($\mathbf{x}$) - μ or median($\mathbf{x}$) - μ, the p-values suggest no statistical differences.

The one-sample t test gives single values; while, one-sample bootstrap tests with various choices on test statistics give different answers. Which test statistics we should use and which answer we put more trust in? In Example 2.3, since the reasons of low speeds are unknown, we practically treat the low speeds as normal data, and still use mean($\mathbf{x}$) - μ. Then, we have evidences to not accept H_0, meaning that there is a speed difference. If we know the low speeds are caused by vehicle malfunctioning or other reasons, we could either eliminate the low speeds at the beginning or use robust test statistics (e.g. trimmed mean or median) to conduct one-sample bootstrap test. In addition, it is recalled that one-sample t test requires the normality assumption. It is unreasonable in practice because the distribution of speeds may not follow Gaussian distributions, which violates the assumption. Overall, we may doubt the conclusion of one-sample t test in Example 2.3.

1.2.4 Summary

The one-sample t test requires that 1)the normality assumption, the null hypothesis under t distributions, and the fixed test statistic. In practice, we may need more flexibility to conduct one-sample hypothesis testing. The one-sample bootstrap test offers us a flexible way to determine whether a value comes from a population with a specific mean. The flexibility of using one-sample bootstrap test includes 1) various choices on test statistics; 2) empirical null hypothesis distributions; 3)various choices on confidence intervals; and 4) no subject to the normality assumption. Unlike that one-sample t test is associated with a closed-form solution, one-sample bootstrap test replies on the resampling with replacement, empirical null hypothesis distributions, and the construction of confidence intervals. Therefore, the accuracy of empirical null hypothesis distributions determines final conclusion. In order to balance between the accurate estimation of empirical null hypothesis distributions and computing, we suggest making $\mathbf{B}$ ranging from 1000 to 10000. In our examples, we took $\mathbf{B}$ as 5000.

1.3 Compute-intensive two-sample test

We use the bootstrap technique to conduct compute-intensive one-sample, which is based on the resampling with replacement. However, it is inappropriate approach to apply the bootstrap to conduct two-sample tests. We use one-sample tests to determine if a value statistically equals a population mean. Unlike one-sample tests, two-sample tests are proposed to determine if two population means equal. Permutation tests and randomization tests are designed to handle two-sample tests, of which are based on the resampling without replacement (a.k.a, shuffling). This section aims to introduce the basic concepts and steps to perform compute-intensive two-sample tests.

1.3.1 Permutation test and randomization test

Let's assume that we collect two samples of data $\mathbf{x_1}$ with N_1 elements and $\mathbf{x_2}$ with N_2 elements. The null and alternative hypotheses are stated as $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 - \mu_2 \neq 0$, respectively. The hypotheses

Table 4: Example of shuffling

		$\mathbf{x_1}$	$\mathbf{x_2}$
origin data		x_1^1, x_1^2	x_2^1, x_2^2, x_2^3
all shuffling	1	x_1^1, x_1^2	x_2^1, x_2^2, x_2^3
	2	x_1^1, x_2^1	x_1^2, x_2^2, x_2^3
	3	x_1^1, x_2^2	x_2^1, x_1^2, x_2^3
	4	x_1^1, x_2^3	x_2^1, x_2^2, x_1^2
	5	x_2^1, x_1^2	x_1^1, x_2^2, x_2^3
	6	x_2^2, x_1^2	x_2^1, x_1^1, x_2^3
	7	x_2^3, x_1^2	x_2^1, x_2^2, x_1^1
	8	x_2^1, x_2^2	x_1^1, x_1^2, x_2^3
	9	x_2^1, x_2^3	x_1^1, x_2^2, x_1^2
	10	x_2^2, x_2^3	x_2^1, x_1^1, x_1^2

indicate that we less care about which mean is greater. Therefore, a nature of test statistics is selected as $|\bar{x_1} - \bar{x_2}|$. The solution of the permutation test for the two-sample problem consists of the following steps:

1. apply the selected test statistic on the two observed samples of data and then it gives θ_{ob};
2. shuffling and reorganize the origin data as two new samples. $\mathbf{x_1}(\mathbf{i})$ and $\mathbf{x_2}(\mathbf{i})$ denote the reorganized data after the *ith* shuffling;
3. apply the select test statistic on $\mathbf{x_1}(\mathbf{i})$ and $\mathbf{x_2}(\mathbf{i})$. The result is denoted as θ_i;
4. after M times of shuffling, we get $\hat{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)$. M can be calculated using Equation 29;
5. p value is calculated as $\frac{\#(\hat{\theta} > \theta_{ob}) + 1}{M+1}$, where, $\#(\cdot)$ is to calculate the size of given conditions.

$$\binom{N_1 + N_2}{N_1} = \binom{N_1 + N_2}{N_2} = \frac{(N_1 + N_2)!}{(N_1)!(N_2)!} \tag{29}$$

Correctly list all of the possible permutations is important to calculate $\hat{\theta}$. Table 4 demonstrates an example given the origin data $\mathbf{x_1}$ with 2 elements and $\mathbf{x_2}$ with 3 elements. The number of possible permutations is 10 including the origin data itself. Evaluating each permutation gives θ_i.

However, enumerating all of the permutations may be exhausted when the sizes of samples go up. For example, M equals 155117520 when each sample has only 15 elements. Randomization tests provide an alternative approach to handle *too many* permutations by randomly selecting from all permutations.

1.3.2 Two-sample randomization test

The procedures of performing randomization tests have no significant differences with the steps of performing permutation tests but the selection of M. Usually we take M from 1000 to 10000 in order to balance the accuracy of the test and the computing. $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 - \mu_2 \neq 0$ are the null and alternative hypotheses.

Example 3.1

Revisit Example 1.3. Traffic professionals collect speed data at daytime and nighttime to determine whether a difference in speed statistically exist. The classical two-sample t test gives the p-value 0.0783, indicating that no significant differences in speed is observed under the significant level 0.05. Let's plug the data into randomization tests and select $|\bar{x_1} - \bar{x_2}|$ to be the test statistic.

After going through the permutation test steps and making M 5000, it gives the p-value 0.0764, which is fairly close to the number generated from the classical t test. Both of the numbers lead to the same conclusion.

Figure 4 shows the empirical distribution of $\hat{\theta}$. Since we take the absolute value of the difference in mean, the minimum θ_i is greater or equal to 0. The solid line represents the θ_{ob} and the filled bars show the area of $\#(\hat{\theta} > \theta_{ob})$.

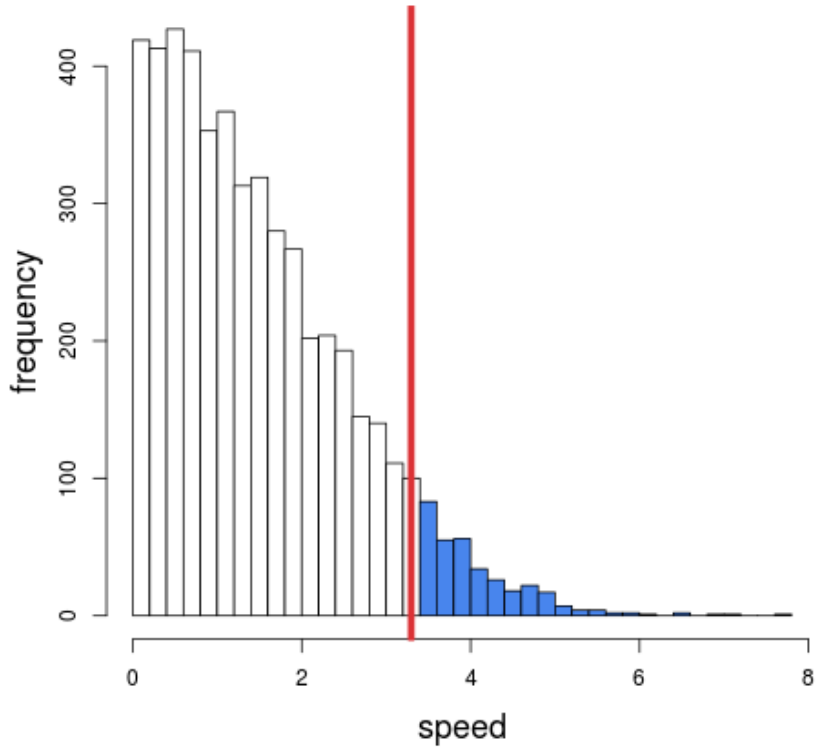


Figure 4: Two-sample randomization test. The solid line is the observed test statistic. Filled bars are the values greater than the observed test statistic

1.3.3 Stratified randomization test

Let's use an example to introduce the concept of *stratified or stratification*.

Example 3.2

Traffic professionals actually collect the first 15 speed data during daytime and 10 speed data at nighttime in January; the rest of the data is collected in February. They have no clue whether *different month* could be an factor to affect the test results. So we help traffic professionals to stratify the speed data by month. *Month* in this example is the stratification.

Table 5 shows the organization of the speed. Instead of shuffling the data together, we only shuffling the data within January or February. Table 6 gives a possibility of the shuffling within the two months. After shuffling within the same month, new samples of speed at daytime and nighttime are produced. Then, we calculate the difference in mean of the new samples.

Note that we still could perform a permutation test. However, Equation 30 gives $M \binom{15+10}{15} * \binom{10+10}{10} = 603923022560$, indicating that the permutation test may not be practice-ready.

Table 5: Stratified speed data

	January	February	average
daytime	45.6, 50.2, 43.7, 43.2, 49.7, 45.5, 45.1, 40.6, 30.9, 43, 42.2, 47.5, 48.1, 39.5, 46.3	41.9, 53.6, 44.6, 53.8, 44.8, 52.4, 46, 44, 46.1, 54.7, 40.5, 40.7, 55.8, 49, 35.9, 53.5, 50.8, 38.2, 42.1, 52.3	44.07
nighttime	44.3, 39.5, 49.3, 28.9, 49.7, 36.8, 32.5, 41, 48.7, 45.1	55.3, 41.4, 34.5, 53.8, 46.3, 27.2, 45.1, 44.3, 36.2, 49.5	41.58

Table 6: A possibility of stratified speed data within month

	January	February	average
daytime	50.2, 43, 30.9, 45.1, 41, 44.3, 49.3, 45.5, 43.2, 48.1, 46.3, 40.6, 47.5, 39.5, 48.7	53.8, 52.4, 55.8, 44.8, 44.6, 53.5, 40.5, 50.8, 53.8, 40.7, 53.6, 55.3, 46, 38.2, 27.2, 52.3, 46.1, 44.3, 41.4, 34.5	45.51
nighttime	45.1, 39.5, 42.2, 49.7, 36.8, 32.5, 49.7, 28.9, 43.7, 45.6	54.7, 49.5, 41.9, 35.9, 42.1, 49, 45.1, 36.2, 46.3, 44	42.92

$$\prod_{s=1}^{S} \binom{n_1^s + n_2^s}{n_1^s} \tag{30}$$

where, n_1^s and n_2^s mean the number of data in the sth stratification of the two sample data.

Figure 5 also shows θ_{ob} and the $\#(\hat{\theta} > \theta_{ob})$ area. The p-value is calculated as 0.0792, which is similar to the case of using non-stratified data 0.0764. Therefore, in Example 3.2, we can draw a conclusion that no significant differences can be observed in speed during daytime and nighttime and *different month* has no significant impact on the speed differences.

1.3.4 Summary

We use the reampling with replacement (bootstrap) to perform one-sample tests; while, the resampling without replacement (shuffling) to perform two-sample tests. Two approaches are beneficial from shuffling: permutation and randomization tests. Randomization tests are special cases of permutation tests. When the number of permutations are too large, we randomly select a subset of permutation, instead of evaluating all possible permutations. Once again, either permutation or randomization tests offer the flexibility to select test statistics, weaken the normality assumption, and estimate test statistic distributions. The test statistic distributions in both examples shows that the distributions do not have to be symmetric or near-symmetric; and p-values are calculated in an intuitive way.

1.4 Compute-intensive ANOVA test

The randomization ANOVA test is also based on the resampling without replacement (a.k.a. shuffling). There are no major differences when performing randomization two-sample and ANOVA test, but the selection of test statistic. The following procedure for the randomization ANOVA is adapted from the basic steps of permutation tests.

1. apply the F-ratio or SS_B on the K samples of data and then it gives θ_{ob}. Selecting F-ratio is mathematically equivalent to selecting SS_B, because the shuffling does not change SS_T;
2. shuffling and reorganize the origin data as K new samples of data. $x_k(i)$ denotes the kth sample in the reorganized data after the ith shuffling;

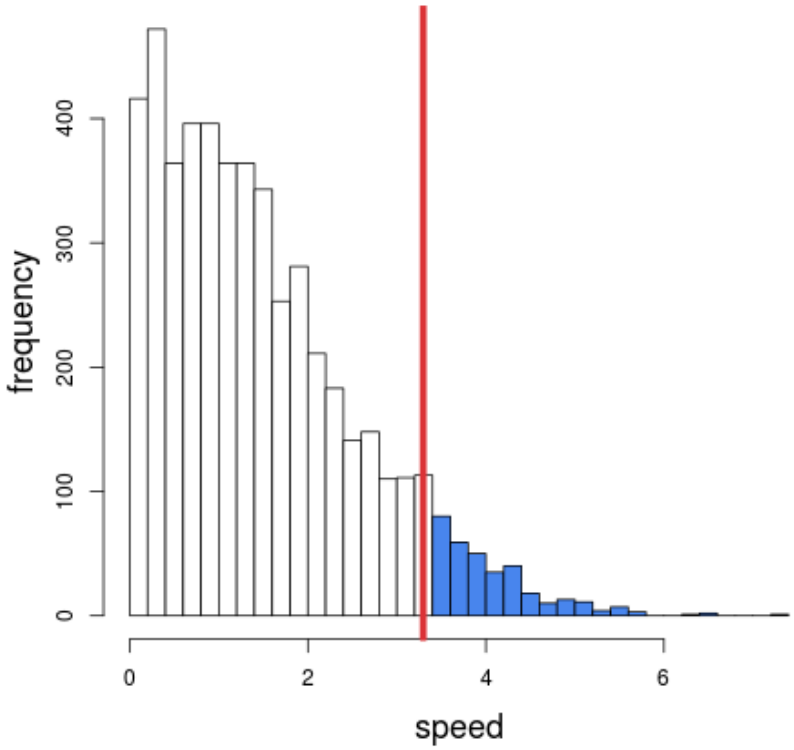


Figure 5: Two-sample stratified randomization test. The solid line is the observed test statistic. Filled bars are the values greater than the observed test statistic

3. apply the select test statistic on $x_1(i), x_2(i), \ldots, x_K(i)$. The result is denoted as θ_i;
4. after M times of shuffling, we get $\hat{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)$. M can range from 1000 to 10000;
5. p value is calculated as $\frac{\#(\hat{\theta} > \theta_{ob}) + 1}{M+1}$, where, $\#(\cdot)$ is to calculate the size of given conditions (Crowley (1992), Phipson and Smyth (2010)).

We herein give the step-based procedures to perform compute-intensive tests, including one-sample, two-sample, and ANOVA. The flow chart-based procedure and calculation steps can be found in the research by Crowley (1992), Noreen (1989), and Efron and Tibshirani (1994).

Example 4.1

Revisit Example 1.4 to determine the existence of the seasonal impact on travel time, and follow the procedures above to calcualte the p-value.

The result shows that the p-value generated from the randomization ANOVA test is 0.14, and the classical ANOVA test gives 0.138. Both p-values suggest that the seasonal impact on travel time is not significant. Figure 6 shows the calculation of the p-value.

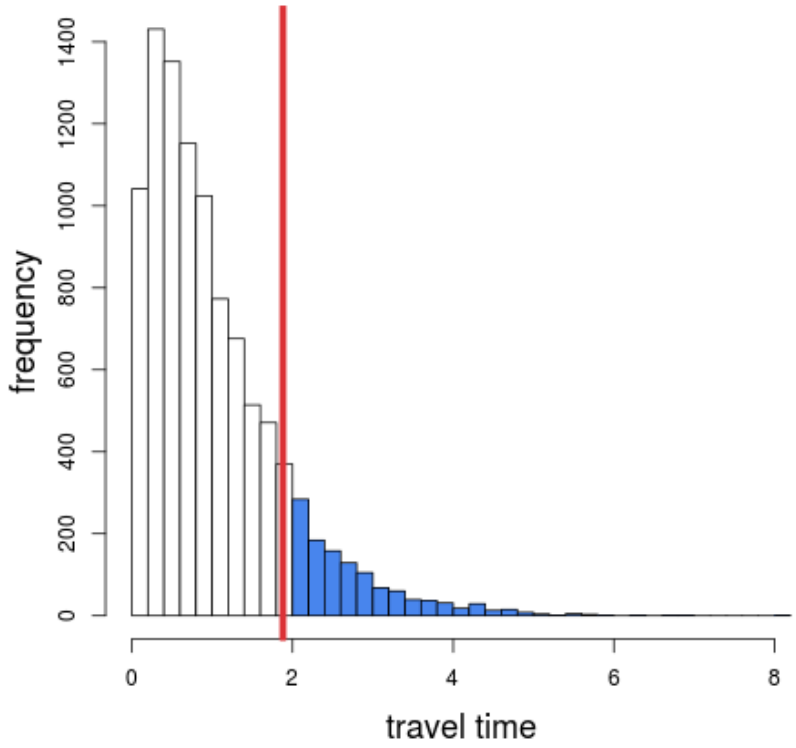


Figure 6: Randomization ANOVA test. The solid line is the observed test statistic. Filled bars are the values greater than the observed test statistic

1.5 Further reading

Both the classical and compute-intensive hypothesis tests have been introduced in this chapter. The following further reading may give a different view of hypothesis testing.

1.5.1 A complete picture of hypothesis testing

In this chapter, we mainly focus on test statistic selection and calculation, null distribution estimation, and the p-value calculation. However, the p-value may not give a complete picture of the result interpretation in hypothesis testing. Confidence intervals concerning about the test statistics may enhance the understanding of the results of hypothesis tests. The effect size is another measure to give a different view of the results. More details of the concepts and applications can be found in the work by Coe (2002), Keselman et al. (2008), and Sullivan and Feinn (2012).

1.5.2 Applications of bootstrap

The bootstrap has been widely used to estimate confidence intervals of a given estimator. The hypothesis testing using the bootstrap is one of the applications. More general applications can be found in Efron and Tibshirani (1994) and Kohavi (1995). and one of the traffic engineering applications can be found in Yang et al. (2017).

1.5.3 Spatial randomness test

We handled one-dimensional sample in this chapter. However, compute-intensive hypothesis testing can be easily adapted in high-demensional data. For example, Clark and Evans (1954) proposed a nearest distance to neighbors based test statistic to determine whether a spatial dataset is randomly scattered. More examples can be in the work by Besag and Diggle (1977) and Crowley (1992).

Reference

Besag, Julian, and Peter J Diggle. 1977. "Simple Monte Carlo Tests for Spatial Pattern." *Applied Statistics*, 327–33.

Clark, Philip J, and Francis C Evans. 1954. "Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations." *Ecology* 35 (4): 445–53.

Coe, Robert. 2002. "It's the effect size, stupid: What effect size is and why it is important." https://www.leeds.ac.uk/educol/documents/00002182.htm.

Crowley, Philip H. 1992. "Resampling Methods for Computation-Intensive Data Analysis in Ecology and Evolution." *Annual Review of Ecology and Systematics* 23: 405–47.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Keselman, H. J., James Algina, Lisa M. Lix, Rand R. Wilcox, and Kathleen N. Deering. 2008. "A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes." *Psychological Methods* 13 (2): 110–29. doi:10.1037/1082-989X.13.2.110.

Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *Appears in the International Joint Conference on Articial Intelligence (IJCAI)* 5: 1–7. doi:10.1067/mod.2000.109031.

Noreen, Eric W. 1989. *Computer-intensive methods for testing hypotheses: an introduction*. New York: Wiley.

Phipson, Belinda, and Gordon K. Smyth. 2010. "Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn." *Stat Appl Genet Mol Biol* 9 (1). doi:10.2202/1544-6115.1585.

Sullivan, Gail M, and Richard Feinn. 2012. "Using Effect Size - or Why the P Value Is Not Enough." *Journal of Graduate Medical Education* 4 (3): 279–82. doi:10.4300/JGME-D-12-00156.1.

Yang, Shu, Chengchuan An, Yao-Jan Wu, and Jingxin Xia. 2017. "Origin-Destination Based Travel Time Reliability." *Transportation Research Record: Journal of the Transportation Research Board* 2643. doi:10.3141/2643-16.