Two sample problems

STAT-S 520

Arturo Valdivia

Modified on April 4, 2023. Subject to change.

Contents

1 Two-Sample Methods

These notes accompanying the methods covered in ISIR (mainly 11.1 but to some extent 10.1. too) as well as SIDS Modern Dive (8.4 and 9.3). Let's start with important questions that help us determine the type of test we need to run:

1. What is the experimental unit? (The experimental units must be independent.)
2. From how many populations were the experimental units drawn? (Remember that the units within each population must be identically distributed.)
   - How many units from were drawn from each population?
   - Is this a 1- or 2-sample problem?
3. How many measurements were taken on each experimental unit? Identify them
4. Define the parameter(s) of interest for this problem: μ for 1-sample problems or Δ for 2-sample problems.

For one-sample location problems, let the random sample be $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{P}$ with $EX_i = \mu$ and $VarX_i = \sigma^2$ for $i = 1, \ldots, n$, where X_i is obtained as a function of the measurement(s) taken on each experimental unit i. The parameter of interest is the population mean μ, but, as explained in ISIR 10.2 and 10.3, other parameters could also be used such as q_2 or θ.

For two-sample location problems:

- Let the first random sample be $X_1, X_2, \ldots, X_{n_1} \overset{iid}{\sim} \mathcal{P}_1$ with $EX_i = \mu_1$ and $VarX_i = \sigma_1^2$ for $i = 1, \ldots, n_1$ and
- Let the second random sample be $Y_1, Y_2, \ldots, Y_{n_2} \overset{iid}{\sim} \mathcal{P}_2$ with $EY_j = \mu_2$ and $VarX_i = \sigma_2^2$ for $j = 1, \ldots, n_2$.
- Each X_i is obtained as a function of the measurement(s) taken on experimental unit i on the first sample, and the first sample mean is given by

  $$\bar{X} = \bar{X}_{n1} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$$

  with $E\bar{X} = \mu$ and $Var\bar{X} = \sigma_1^2/n_1$.
- Y_j is obtained as a function of the measurement(s) taken on each experimental unit j on the second sample, and the second sample mean is given by

  $$\bar{Y} = \bar{Y}_{n_2} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j$$

  with $E\bar{Y} = \mu_2$ and $Var\bar{Y} = \sigma_1^2/n_2$.
- The parameter of interest is usually the *difference* in population means:

  $$\Delta = \mu_1 - \mu_2$$

  Note that it is not important which sample you call the X's and which you call the Y's, as long as you are consistent throughout the analysis (you should be specially careful when stating the inequalities needed for the hypotheses).
- The estimator of the difference of means is

  $$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{X} - \bar{Y}.$$

  If

  $$\bar{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \bar{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

and using the properties of the expected value and variance, the distribution of $\hat{\Delta}$ is

$$\hat{\Delta} \sim \mathcal{N}\left(\Delta, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

5. Do you need to perform a test of significance? If so, what are the appropriate null and alternative hypotheses? In a one-sample location problem, the hypotheses should be statements about μ (or θ.) In a two-sample location problem, the hypotheses should be statements about Δ.

2 Welch's Two-Sample t-test

When we have two samples from approximately normal populations, there are three options for significance tests concerning the difference in means.

- Use the normal distribution for $\hat{\Delta}$, which is justifiable if you know both population standard deviations. This seldom happens in practice, with the exception of the special case where the problem can be constructed for Bernoulli/Binomial random variables.
- When we have two i.i.d. samples from two independent normal populations, we can use the **Welch's two-sample t-test.**
  - When, in addition, we know that $\sigma_1^2 = \sigma_2^2$ (equal variances), we could use the **Student's two-sample t-test**.
  - Given that we almost never know whether population variances are equal and when they are not the Student's test could potential give bad results, we always prefer to use the Welch's t-test.

The Welch's two-sample t-test statistic is

$$T_W = \frac{\hat{\Delta} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where Δ_0 is the hypothesized value of the difference of means under the null hypothesis. The distribution of T_W approximates a t-distribution with

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\left(\sigma_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(\sigma_2^2/n_2\right)^2}{n_2 - 1}}$$

degrees of freedom and can be estimated by

$$\hat{\nu} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

Finally, using Welch's approximation, a $(1 - \alpha) \times 100\%$ confidence interval for Δ is

$$\hat{\Delta} \pm q_t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where $q_t =$ `qt(1-`$\alpha/2, \hat{\nu}$`)`. For a detailed construction and explanation of the tests of significance and confidence interval, please read ISIR pp. 278-280.

2.1 Assumptions of Welch's t-test

Welch's t-test assumes two independent samples from normally distributed populations. However, the test is fairly *robust* to minor violations of its assumptions: that is, it's still acceptable (in terms of Type I and II error estimated probabilities) if the underlying distributions are not quite normal. In practice, you can usually get away with Welch's test as long as your samples are not small and they are reasonably symmetric with no gross outliers. On the other hand, if you have strongly skewed data or bad outliers and the sample size are not large, you should consider a transformation, simulation or bootstrapping methods, or a nonparametric test.

As usual, the larger the samples the better, both in terms of robustness to assumptions and (usually more importantly) in terms of power.

3 Developing the 2-sample Welch's t-test based on case: "Etruscan skulls"

3.1 The case study

Were the skull sizes of ancient Etruscans different from the skull sizes of modern Italians? The problem is presented in ISIR pp. 290-294; the data can be obtained from the book's website. Before we answer the five basic questions, we'll take a look at the data.

In the data set as posted, the first 84 numbers are the breadths (in mm) of skulls of ancient Etruscan men, while the remaining 70 numbers are breadths of a sample of skulls of ancient Italian men. (I don't know if the samples are random – in particular, it's hard to imagine how one would take a truly random sample of skulls of long-dead Etruscans – but we'll assume there were no systematic biases in the data collection.)

```
vec.skulls = scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/skulls.dat")
etruscan = vec.skulls[1:84]
italian = vec.skulls[85:154]
```

3.1.1 Exploring the data

Have a look at the numerical summaries:

```
summary(etruscan)
```

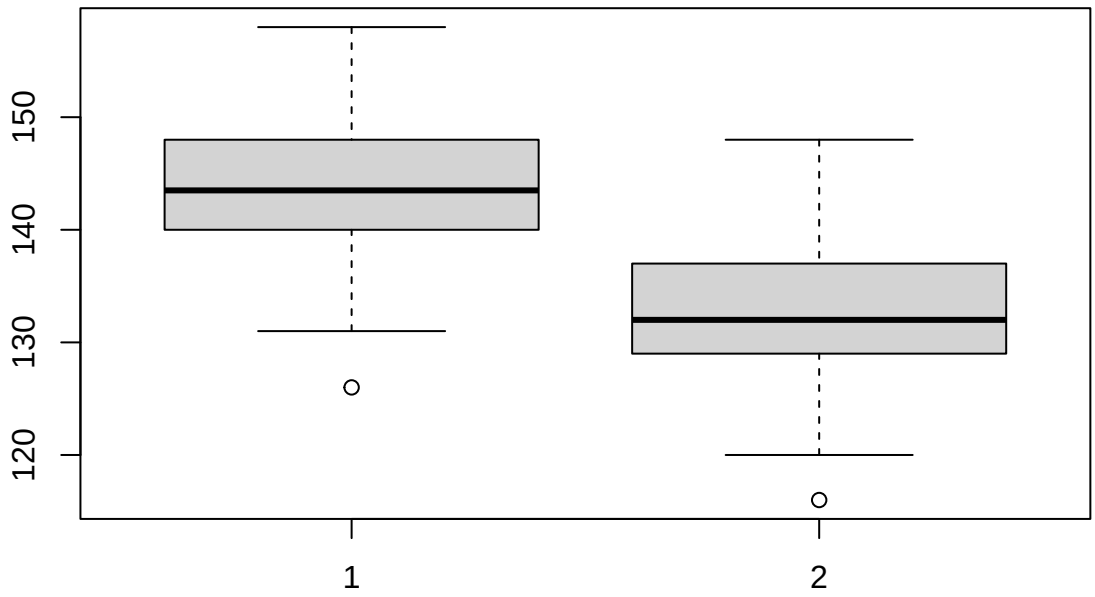
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   140.0   143.5   143.8   148.0   158.0
```

```
summary(italian)
```

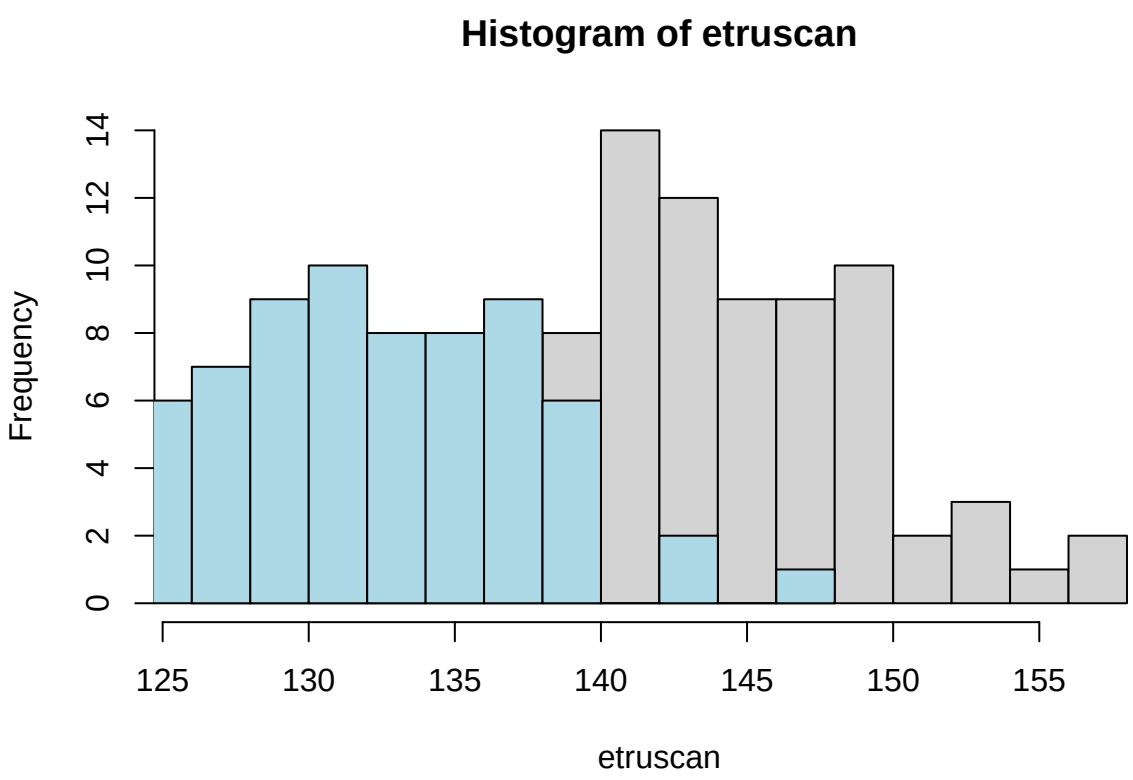
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   116.0   129.0   132.0   132.4   136.8   148.0
```

And draw some pictures:

```
boxplot(etruscan, italian)
```




```
hist(etruscan, breaks = 20)
hist(italian, breaks = 20, col="lightblue", add=T)
```



It looks like the Italian (blue) distribution is shifted to the left compared to the Etruscan (grey) distribution – and the sample sizes are reasonable. Still, there might still be some doubt in your mind as to whether a difference of this size could be explained by chance, so let's do a significance test.

Now let's answer our basic questions:

- The experimental unit is a skull.
- The skulls are sampled from two populations: ancient Etruscans and modern Italians.
- One measurement is taken on each skull – the breadth, in millimeters.
- Let X_i be the breadth of the ith Etruscan skull, and Y_j be the breadth of the jth Italian skull. The parameter of interest is $\Delta = \mu_1 - \mu_2$. In the next section, we recap the theory needed to understand why Δ is the parameter of interest.

3.2 Theoretical Setting

Let's review the theory described above in the context of this example. Feel free to skip to the next section to continue working on the solution.

The assumptions we make for the theory-based approach are that

$$X_1, \ldots, X_{84} \overset{iid}{\sim} \mathcal{P}_1$$

with $EX_i = \mu_1$ for $i = 1, \ldots, 84$ and

$$Y_1, \ldots, Y_{70} \overset{iid}{\sim} \mathcal{P}_2$$

with $EY_j = \mu_2$ for $j = 1, \ldots, 70$. Based on the plots obtained, the distribution of X_i and Y_j are somewhat symmetric and without obvious outliers. Since the samples are large enough, the CLT makes the distributions

of the sample means near normal; namely

$$\bar{X}_{84} \dot{\sim} \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{84}\right) \qquad \text{and} \qquad \bar{Y}_{70} \dot{\sim} \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{70}\right).$$

Moreover, we can construct the difference of sample means, as a random variable; namely

$$\hat{\Delta} = \bar{X}_{84} - \bar{Y}_{70} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{84} + \frac{\sigma_2^2}{70}\right).$$

Recall that Δ is defined as

$$\begin{aligned} \Delta &= E(\bar{X}_{84} - \bar{Y}_{70}) \\ &= E\bar{X}_{84} - E\bar{Y}_{70} \\ &= \mu_1 - \mu_2 \end{aligned} \tag{1}$$

3.3 Test of Significance

3.3.1 Hypotheses

There was no direction to the test given, so a two-tailed test seems appropriate here

$$\begin{aligned} H_0 &: \Delta = 0 \\ H_1 &: \Delta \neq 0 \end{aligned}$$

Note this is the same as testing

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

3.4 Theory-based approach: using the Welch's two-sample T-test

Our point estimate of Δ is just the difference in sample means:

```
mean(etruscan) - mean(italian)
```

```
## [1] 11.33095
```

We call this $\hat{\Delta}$ ("Delta hat" to indicate that this is an estimate of the true population value, Δ.)

```
Delta.hat = mean(etruscan) - mean(italian)
```

The estimated standard deviation of $\hat{\Delta}$, often known as the **standard error** of $\hat{\Delta}$, is:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
se = sqrt(var(etruscan)/84 + var(italian)/70)
```

Assuming a null hypothesis of no difference, the Welch's t-statistic is $\hat{\Delta}$ divided by its standard deviation:

```
t.Welch = (Delta.hat - 0)/se
print(t.Welch)
```

```
## [1] 11.96595
```

The approximate degrees of freedom are:

$$\hat{\nu} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

```
nu = (var(etruscan)/84+var(italian)/70)^2/
  ((var(etruscan)/84)^2/83+(var(italian)/70)^2/69)
print(nu)
```

```
## [1] 148.8193
```

Typically, your approximated degrees of freedom, $\hat{\nu}$ won't be a whole number. This won't be a problem since R can handle decimal degrees of freedom.

If the null hypothesis is true, then Welch's t-statistic has an approximate t-distribution with $\hat{\nu}$ degrees of freedom. For a two-tailed test, the p-value is thus:

```
P.value = 2*(1-pt(abs(t.Welch), df=nu))
P.value
```

```
## [1] 0
```

The P-value is basically zero and we have strong evidence to conclude that there is a difference on average Etruscan and Italian skulls' breath.

3.5 Simulation-based Approach (Bootstrapping) for the Comparison of Two-Sample Means

We first reformat the vector of skulls' measurements as a data frame:

```
vec.skulls = scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/skulls.dat")
group = c(rep("etruscan",84), rep("italian",70))
data.skulls = data.frame(group, breath = vec.skulls)
```

We can now create the bootstrap samples of the difference in means:

```
library(infer)
null_dist <- data.skulls %>%
  specify(breath ~ group) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("etruscan", "italian"))
```

We get again the estimate difference based on the original samples

```
delta_hat <- data.skulls %>%
  specify(breath ~ group) %>%
  calculate(stat = "diff in means", order = c("etruscan", "italian"))
delta_hat
```

```
## Response: breath (numeric)
## Explanatory: group (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  11.3
```

and finally obtain the P-value, that is zero as in the theory-based approach.

```
null_dist %>%
  get_p_value(obs_stat = delta_hat , direction = "two-sided")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

4 Developing the 2-sample confidence intervals based on case: "Etruscan skulls"

4.1 Theory-based approach

Let's obtain a 95% confidence interval. We start from $\hat{\Delta}$ and go up and down by q standard errors (where the standard error is the same one we found earlier), where q is determined using a t-distribution with the appropriate number of degrees of freedom.

```
alpha = 1-0.95
q = qt(1 - alpha/2, df=nu)
lower = Delta.hat - q*se
upper = Delta.hat + q*se
lower
```

```
## [1] 9.459782
```

```
upper
```

```
## [1] 13.20212
```

We are 95% confident that the ancient Etruscan average skull breadth was 9–13 mm more than the modern Italian. In addition, if the confidence interval is used as a complement of the prior test of significance, observe that 0 is not part of the interval, supporting the conclusion that the average skull breaths were different.

4.2 Simulation-Based (Bootstrapping) Approach

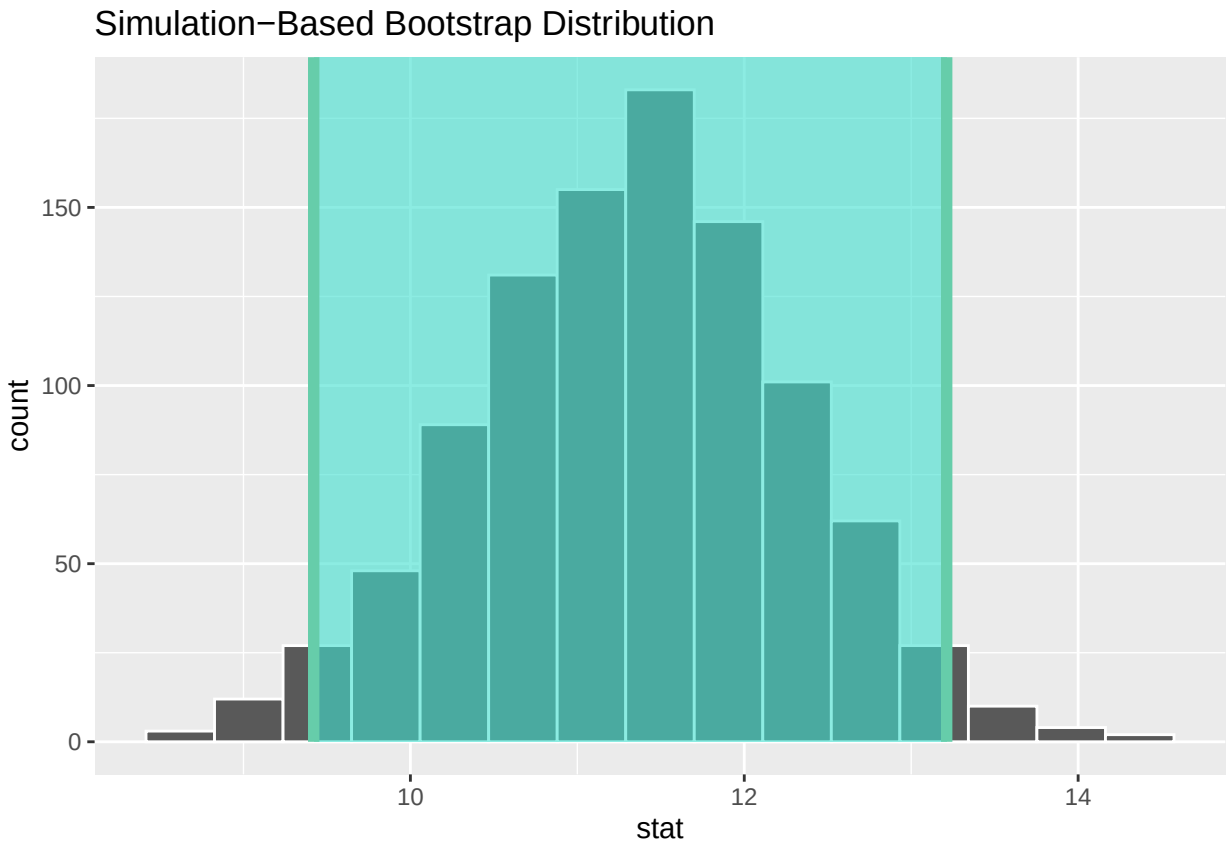
Let's now obtain the 95% confidence interval using the simulation-based approach:

```
boot_dist <- data.skulls %>%
  specify(breath ~ group) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("etruscan", "italian"))

percentile_ci <- get_ci(boot_dist, level = 0.95)
percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     9.42     13.2
```

```
visualize(boot_dist) +
  shade_confidence_interval(endpoints = percentile_ci)
```



The confidence interval is very similar to one obtained using the theory-based approach.

5 The function `t.test()` in R

The function `t.test()` allows to perform many one- and two-sample T-tests, including the Welch's:

```
Welch.t = t.test(x = etruscan, y = italian, alternative = "two.sided", conf.level = 0.95)
Welch.t
```

```
##
##  Welch Two Sample t-test
##
## data:  etruscan and italian
## t = 11.966, df = 148.82, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.459782 13.202123
## sample estimates:
## mean of x mean of y
##  143.7738  132.4429
```

or, if specific information is required:

```
Welch.t$p.value
```

```
## [1] 1.530613e-23
```

```
Welch.t$conf.int
```

```
## [1]  9.459782 13.202123
## attr(,"conf.level")
## [1] 0.95
```