

Lecture 24: Comparing two population means

Chapter 18

Instructors: Tomer Altman and Alan Hubbard

October 25, 2024

Recap

- So far we've learned the z-test and the t-test that apply when the variable of interest is continuous
- These are **one sample** tests, meaning we have one variable of interest, and we take one sample. We're interested in knowing whether the one sample differs from some null hypothesized value ($H_0 : \mu = 68mm$)

Next up: comparing two population means

- In Chapter 18, we are interested in comparing two samples from two populations. We want to shed light on the question: Do these two samples have the same underlying mean, or are their means different? Another way of phrasing this is, "is the difference between the means equal to 0 or not?"
- For this question, our null and two-sided alternative hypotheses are:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

- You can also write them this way, by rearranging the equation:

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

Comparing two population means

In public health, we are often most interested in questions that compare two groups (exposed vs. unexposed, treatment vs. placebo, etc)

- Run a randomized controlled trial where we compare a treated subgroup to a placebo group. Do their mean health outcomes differ?
- Conduct an observational study where we have exposed and unexposed individuals. Do their mean health outcomes differ?

In this chapter, we talk about tests to compare the difference in two continuous means, represented by μ_1 and μ_2

Comparing two samples, graphically

Procedure to compare two samples graphically

- Make two histograms, one for each sample
- Compare their shapes, centers (means or medians) and spreads (standard deviations)
- Or, make two box plots and compare their medians and IQRs

Conditions for inference comparing two means

- We have two SRSs, representing two distinct populations.
 - The samples are **independent**. That is, the individuals in one sample are unrelated to the individuals in the other sample.
 - We measure the same quantitative variable for both samples.
- Both populations are **Normally distributed**. The means and standard deviations of the populations are unknown. **In practice, it is enough that the distributions have similar shapes, sample size is “big enough”, and that the data have no strong outliers.**

Notation

Notation for the population parameters:

Population	Variable	Population mean	Population SD
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

Notation

Notation for the sample statistics:

Population	Sample size	Sample mean	Sample SD
1	n_1	\bar{x}_1	s_1
2	n_2	\bar{x}_2	s_2

To perform inference about the difference between the means ($\mu_1 - \mu_2$), we first estimate the difference $\bar{x}_1 - \bar{x}_2$ between the sample means.

Two-sample t test

- With one-sample tests, we have one \bar{x} and we sketch the sampling distribution for \bar{x} . It is centered at μ with standard error σ/\sqrt{n}
- With two-samples, we have two sample averages \bar{x}_1 and \bar{x}_2 .
- What do we do?

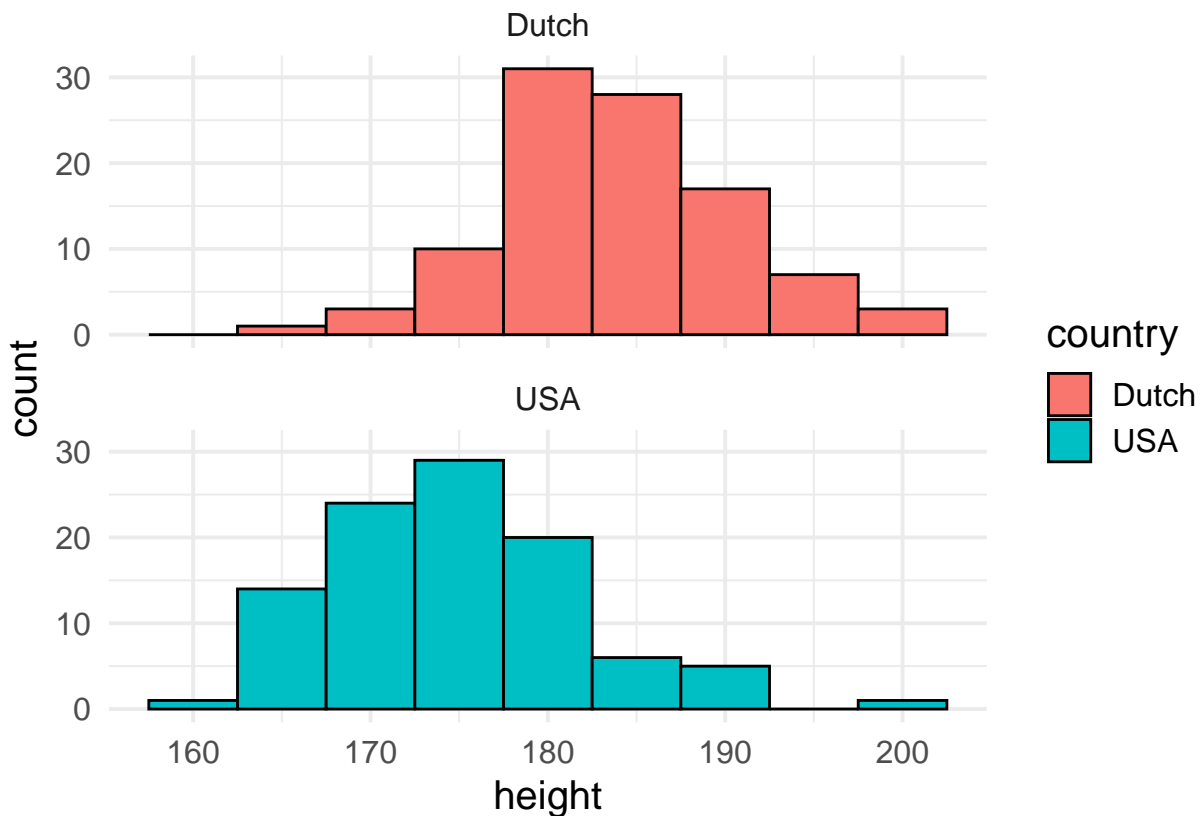
Example with adolescent heights from two different countries

Are the heights of US and Dutch-born men different?

```
## # A tibble: 2 x 4
##   country sample_mean sample_sd length
##   <chr>      <dbl>      <dbl> <int>
## 1 Dutch        184.        6.75   100
## 2 USA          175.        7.16   100
```

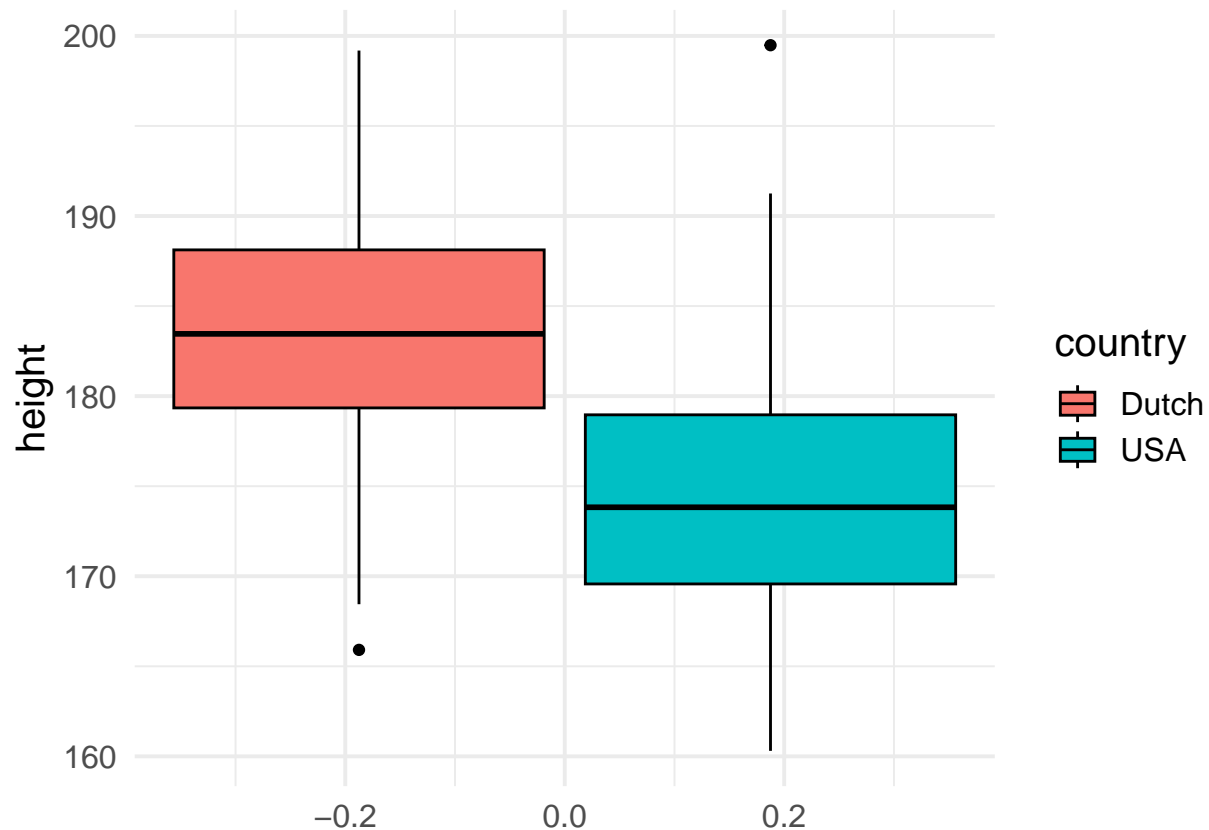
Compare the histograms

```
ggplot(height_data, aes(x = height)) +
  geom_histogram(aes(fill = country), binwidth = 5, col = "black") +
  theme_minimal(base_size = 15) +
  facet_wrap(~country, nrow = 2)
```



Compare the boxplots

```
ggplot(height_data, aes(y = height)) +
  geom_boxplot(aes(fill = country), col = "black") +
  theme_minimal(base_size = 15)
```



Example with adolescent heights from two different populations

Are the heights of US and Dutch-born men different?

$\bar{x}_{USA} = 174.7042$ and $\bar{x}_D = 183.6690$ so the difference is $\bar{x}_D - \bar{x}_{USA} = 8.9648$

What would happen if we took another set of two samples of USA and Dutch-born men? We would expect these sample means to change. We could draw an approximate **sampling distribution for the difference** between these two means.

The sampling distribution of $\bar{x}_1 - \bar{x}_2$

The distribution of the difference between two independent random variables has a mean equal to the difference of their respective means and a variance equal to the sum of their respective variances. That is:

- The mean of the sampling distribution for $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$.
- The variance of the sampling distribution is: $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- The standard deviation of the sampling distribution is: $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Our *estimate* of the standard deviation of the sampling distribution is: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Recall the t-test

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

We need to generalize this by replacing each piece in the z-test by the calculations on the previous slide:

The **two-sample** t-test is therefore:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The two-sample t statistic has approximately a t distribution. The approximation is accurate when both sample sizes are greater than or equal to 5.

The degrees of freedom (df) for the two-sample t-test...

is bananas. Here is the formula:

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{s_2^2}{n_2})^2}$$

Where:

- s_1^2 is the sample variance for sample #1
- s_2^2 is the sample variance for sample #2
- n_1 is the sample size for sample #1
- n_2 is the sample size for sample #2

Note, that if sample size is large enough, the test statistic will follow a $N(0, 1)$ distribution and this calculation becomes much less important.

Confidence intervals for the two-sample t-test

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where t^* is the critical value with area C between $-t^*$ and t^* under the t density curve with the appropriate degrees of freedom.

Hypothesis testing when you have two samples

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

Here we consider the two-sided alternative hypothesis, but you could be interested in the one-sided alternative in either direction: $H_A : \mu_1 - \mu_2 > 0$

Obtain the **two-sample t-test** statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under the two-sided alternative:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the **p-value** is the probability, assuming H_0 is true, of getting a test statistic t that we saw or a more extreme value. We compute the p-value as the corresponding area under the t distribution with the appropriate degrees of freedom.

Example, continued

Let R do the work for you:

```
t.test(height_data_wide %>% pull(usa),
       height_data_wide %>% pull(dutch),
       alternative = "two.sided")

##
##  Welch Two Sample t-test
##
## data:  height_data_wide %>% pull(usa) and height_data_wide %>% pull(dutch)
## t = -9.1103, df = 197.35, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.905418  -7.024282
## sample estimates:
## mean of x mean of y
##  174.7042  183.6690
```

Note that `t.test` gives you both the t-test results (including the t-statistic (called “t” in the output), df, and p-value), as well as the 95% CI. We got both because we performed a two-sided test.

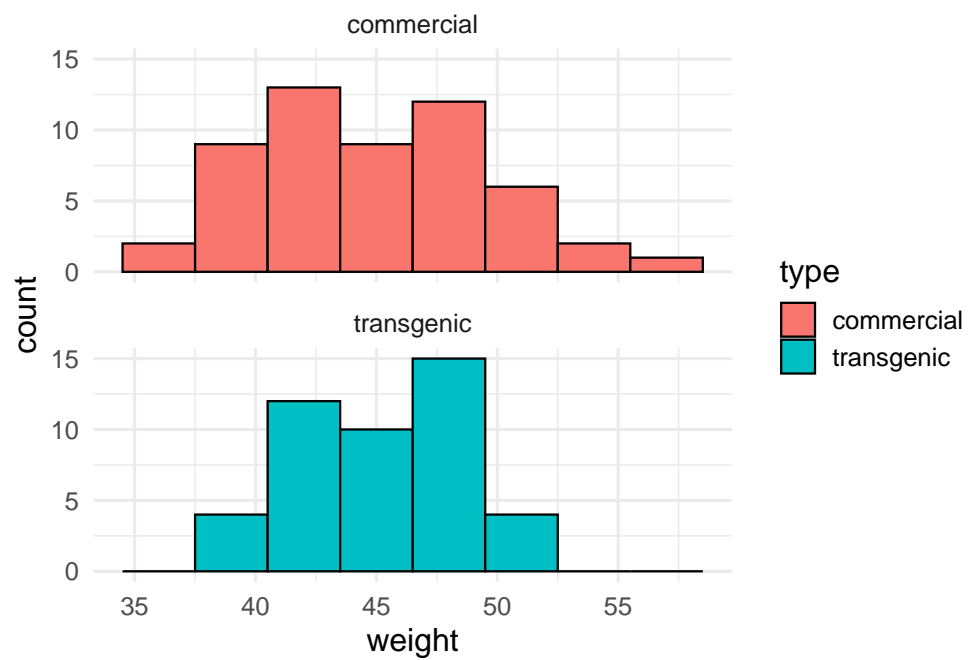
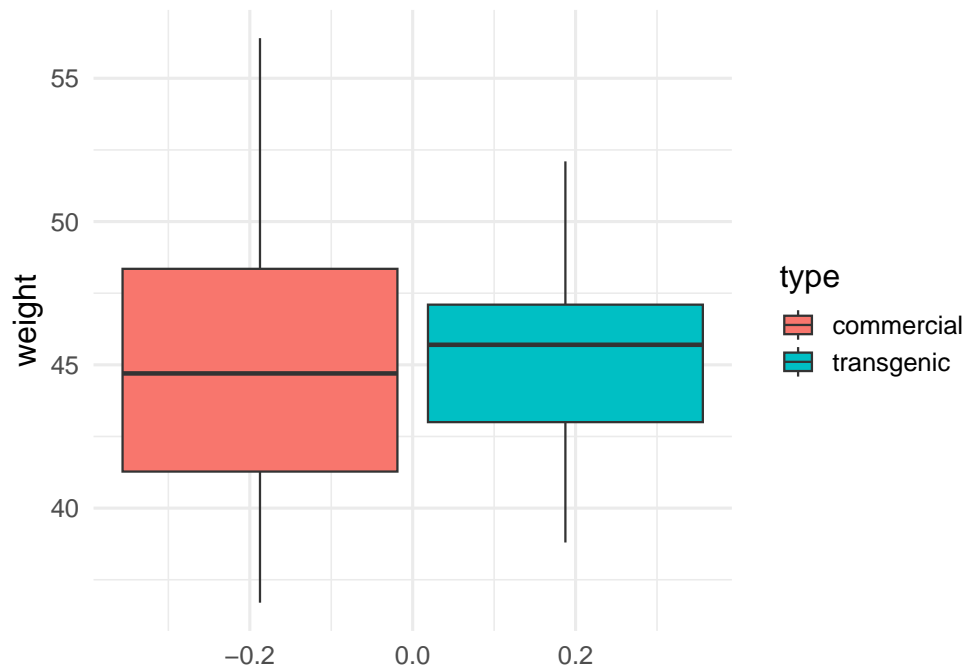
Example 2: Transgenic chickens

Infection of chickens with the avian flu is a threat to both poultry production and human health. A research team created transgenic chickens resistant to avian flu infection. Could the modification affect the chicken in other ways? The researchers compared the hatching weights (in grams) of 45 transgenic chickens and 54 independently selected commercial chickens of the same breed.

```
transgenic <- c(38.8, 39.0, 39.7, 40.0, 40.8, 40.9, 41.0, 41.0, 41.0, 42.5, 42.6, 43.0,
               43.0, 43.4, 43.5, 43.5, 43.8, 44.4, 44.7, 44.7, 44.7, 45.3, 45.7, 45.8,
               46.4, 46.5, 46.6, 46.7, 46.7, 46.8, 46.9, 47.1, 47.1, 47.1, 47.3, 47.6,
               47.7, 48.1, 48.3, 49.3, 49.3, 49.8, 50.3, 50.9, 52.1)

commercial <- c(36.7, 37.1, 38.9, 39.5, 39.5, 39.8, 40.0, 40.2, 40.3, 40.5, 40.5, 40.7,
               41.1, 41.2, 41.5, 41.5, 41.6, 41.6, 41.7, 42.4, 43.1, 43.3, 43.3, 43.4,
               43.7, 44.1, 44.2, 45.2, 45.3, 45.4, 46.0, 46.1, 46.4, 46.6, 46.6, 46.9,
               47.3, 47.5, 48.1, 48.2, 48.4, 48.6, 49.0, 49.1, 49.3, 49.6, 50.1, 50.2,
               50.4, 50.6, 52.2, 53.0, 55.5, 56.4)

chicken_data <- data.frame(weight = c(transgenic, commercial),
                           type = c(rep("transgenic", 45), rep("commercial", 54)))
```



Estimate the size of the difference between the two means

```
means <- chicken_data %>%
  group_by(type) %>%
  summarise(mean_weight = mean(weight))
```

means

```
## # A tibble: 2 x 2
##   type      mean_weight
##   <chr>      <dbl>
## 1 commercial    45.0
## 2 transgenic    45.1
```

```
diff_means <- means[1, 2] - means[2, 2]
diff_means
```

```
##   mean_weight
## 1   -0.153333
```

The estimated mean difference is -0.153 grams.

Estimate the standard error

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

```
chicken_stats <- chicken_data %>%
  group_by(type) %>%
  summarise(mean_weight = mean(weight),
            sd_weight = sd(weight),
            n = length(weight))
```

Use the output to calculate the SE:

$$SE = \sqrt{\frac{4.568872^2}{54} + \frac{3.320836^2}{45}} = 0.7947528$$

Calculate the t-statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{(44.98889 - 45.14222) - (0)}{0.7947528} = -0.1929279$$

What is the chance of observing the t-statistic -0.193 on the t-distribution with the appropriate degrees of freedom?

To answer this, we would need to calculate the degrees of freedom using the complicated formula from a few slides previous. We won't do this. Instead, we will ask R to do the test for us (and verify that our calculated t-statistic matches R's test)

t.test in R

Pay attention to the arguments specified by `t.test`. Below, the first argument is the weight data for the commercial chickens and the second argument is the weight data for the transgenic chickens.

```
commercial_weight <- chicken_data %>% filter(type == "commercial") %>% pull(weight)
transgenic_weight <- chicken_data %>% filter(type == "transgenic") %>% pull(weight)

t.test(commercial_weight, transgenic_weight, alternative = "two.sided")
```



```
##
## Welch Two Sample t-test
##
## data: commercial_weight and transgenic_weight
## t = -0.19293, df = 95.344, p-value = 0.8474
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.731044 1.424377
## sample estimates:
## mean of x mean of y
## 44.98889 45.14222
```

What happens if you flip the order?

- If you flip the order, the test statistic flips signs (from negative to positive or vice versa) as does the values for the CI. The p-value and df stays the same.

Robustness of the two-sample t-test

- When the sizes of the two samples are equal and the two populations being compared have similar shapes, the two-sample t-test will work well for sample sizes as small as $n_1 = n_2 = 5$
 - Though still requires underlying random variable is normal within the two groups if sample size is relatively small in either group.
- When the two populations have different shapes, larger samples are needed (e.g., one skewed left and the other skewed right).
 - This is because the difference of averages (the estimator) becomes normally distributed even if the underlying random variable in both groups is not normal.