# Two-Sample Tests

*Chapter 5, Lab 1: Solutions*

*OpenIntro Biostatistics*

**Topics**

- Distinguishing between paired and independent data
- Two-sample test for paired data
- Two-sample test for independent group data

This lab extends inference for means in the one-sample setting to the two-sample setting. Two-sample data can consist of two samples in which the observations are paired in some way or in which the observations are drawn from two independent groups. In the paired context, inference is made about the population mean of the differences between observations; in the independent context, inference is made about the difference in population means between the two groups.

The material in this lab corresponds to Sections 5.2 - 5.3 of *OpenIntro Biostatistics*.

## Distinguishing between paired and independent data

1. The following scenarios describe examples of two-sample data. Identify whether the data are paired (i.e., an observation in one group can be paired to an observation in the other) or independent.

   a) A multimedia program designed to improve dietary behavior among low-income women was evaluated by comparing women who were randomly assigned to intervention and control groups. The intervention was a 30-minute session in a computer kiosk in the Food Stamp office. About 2 months after the program, the participants took a knowledge test about healthy eating practices. The test scores will be used as an outcome in an analysis assessing the efficacy of the multimedia program.

   The data are independent. The women in the two groups are not paired; the test will compare mean test score for women in the intervention group (who took the multimedia program) versus mean test score for the women in the control group.

   b) An investigator is studying standardized IQ test scores for third grade students in Massachusetts to see if birth order is associated with test score. He has collected test score data for 25 sets of siblings (first-born versus second-born).

   The data are paired by sibling, where the difference in test score can be calculated for each sibling pair. The hypothesis test assesses whether mean difference in scores is significantly different from 0.

   c) Researchers are interested in analyzing the relationship between oral contraceptive use and blood pressure in women.

i. A group of women who are not currently oral contraceptive users are identified and their blood pressure is measured. One year later, the women who have become oral contraceptive users are identified; these women are selected as the study population and their blood pressure is measured a second time. The researchers will compare the initial blood pressure and the follow-up blood pressure of the women in the study population.

The data are paired. In this setting, each women has two blood pressure measurements; one at the initial identification stage when not using oral contraceptives, and one a year later, when using oral contraceptives. The difference in blood pressure can be calculated for each woman.

ii. A group of oral contraceptive users and a group of non-users are identified, and their blood pressure is measured. The researchers will compare blood pressure between the users and non-users.

The data are independent. This test will compare the mean blood pressure in the group of women who are oral contraceptive users to the mean blood pressure in the group of women who are non-users.

**Two-sample test for paired data**

*Did a new wetsuit design allow for increased swim velocities during the 2000 Olympics?*

Wetsuits are commonly used in the swimming stage of triathlons; in addition to providing thermal insulation against cold water, wetsuits are also thought to increase swimming speed. In 2008, de Lucas and co-authors conducted a study to assess the effect of wetsuits on swim velocity. In this study, 12 competitive swimmers were asked to swim 1,500 meters at maximal velocity, once wearing a wetsuit and once wearing a standard swimsuit. The order of wetsuit versus swimsuit was randomized for each swimmer.

The mean velocity (m/s) for each 1500m swim is recorded in the `swim` dataset in `oibiostat`.

2. Load and view the data. There are two velocity values for each swimmer: one for the wetsuit trial and one for the swimsuit trial.

```
#load the data
library(oibiostat)
data("swim")
```

a) For swimmer 1, what is the difference between velocity measured during the wetsuit trial and velocity measured during the swimsuit trial?

The difference between velocity measured during the wetsuit trial and velocity measured during the swimsuit trial for swimmer 1 is 0.08 m/s.

```
swim$velocity.diff[1]
```

```
## [1] 0.08
```

b) How were the values stored in the `velocity.diff` variable calculated? What does a

positive value for `velocity.diff` imply, versus a negative value?[1]

The values are calculated by subtracting swimsuit velocity from wetsuit velocity. Thus, a positive value implies that the velocity was higher in the wetsuit trial; a negative value implies that the velocity was lower in the wetsuit trial.

c) Calculate $\overline{d}$, the mean of the observed differences.

The mean of the observed differences is 0.0625.

```
#calculate d.bar
d.bar = mean(swim$velocity.diff)
d.bar
```

```
## [1] 0.0775
```

3. Conduct a hypothesis test to address the question of interest. Let $\alpha = 0.05$.

a) Suppose the parameter $\delta$ is the population mean of the differences in velocities during a 1500m swim if all competitive swimmers recorded swim velocities with both suit types. State the null and alternative hypotheses.

The null hypothesis is that there is no difference in mean swim velocities when swimming with a wetsuit versus aswim suit, $H_0 : \delta = 0$. The alternative hypothesis is that there is a difference in mean swim velocities when swimming with a wetsuit versus a swimsuit, $H_A : \delta \neq 0$.

b) Calculate the test statistic, where $\overline{d}$ is the mean of the differences, $s_d$ is the standard deviation of the differences, and $n$ is the number of differences (i.e., number of pairs). Note how the formula for the test statistic is identical to the one introduced in the previous chapter; a paired $t$-test is essentially a one-sample test of difference values.

$$t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{n}} = \frac{0.0625 - 0}{0.0585/\sqrt{12}} = 3.70$$

```
#use r as a calculator
d.bar = mean(swim$velocity.diff)
delta.0 = 0
s.d = sd(swim$velocity.diff)
n = length(swim$velocity.diff)

t.stat = (d.bar - delta.0)/(s.d/sqrt(n))
t.stat
```

```
## [1] 12.31815
```

c) Calculate the $p$-value from a $t$-distribution with degrees of freedom $n - 1$.

The $p$-value is $P(T \geq |3.70|) = 2 \times P(T \geq 3.70) = 0.00349$.

```
#calculate the p-value
2*pt(t.stat, df = n - 1, lower.tail = FALSE)
```

[1] Note: there were no negative values in these data.

```
## [1] 8.885414e-08
```

d) Draw a conclusion.

Since $p$ is less than $\alpha = 0.05$, there is sufficient evidence to reject the null hypothesis and accept the alternative. The data support the claim that the wetsuits changed swim velocity in a 1500m swim. The observed average increase of 0.0625 m/s for swimmers wearing wetsuits is significantly different than the null hypothesis of no difference, and suggests that swim velocities are higher when swimmers wear wetsuits instead of swimsuits.

4. Calculate a 95% confidence interval. Interpret the interval in the context of the data.

$$\overline{d} \pm t^\star \times \frac{s_d}{\sqrt{n}}$$

$$0.0625 \pm 2.20 \times \frac{0.0585}{\sqrt{12}}$$

$$(0.0253, 0.0997) \text{ m/s}$$

With 95% confidence, the population mean difference in swim velocities is in the interval (0.0253, 0.0997) m/s. The interval does not include 0 (no change), which is consistent with the result of the hypothesis test.

```
#use r as a calculator
d.bar = mean(swim$velocity.diff)
s.d = sd(swim$velocity.diff)
n = length(swim$velocity.diff)
t.star = qt(0.975, df = n - 1)

m = t.star * (s.d/sqrt(n))
d.bar - m; d.bar + m
```

```
## [1] 0.06365244
```

```
## [1] 0.09134756
```

5. Verify the answers from Questions 3 and 4 using `t.test()`.

```
#two-sample syntax
t.test(swim$wet.suit.velocity, swim$swim.suit.velocity,
   alternative = "two.sided", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  swim$wet.suit.velocity and swim$swim.suit.velocity
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
```

```
## sample estimates:
## mean of the differences
##                  0.0775
```

```
#alternatively, one-sample syntax
t.test(swim$velocity.diff, mu = 0, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  swim$velocity.diff
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of x
##    0.0775
```

**Two-sample test for independent group data**

*Does change in non-dominant arm strength after resistance training differ between men and women?*

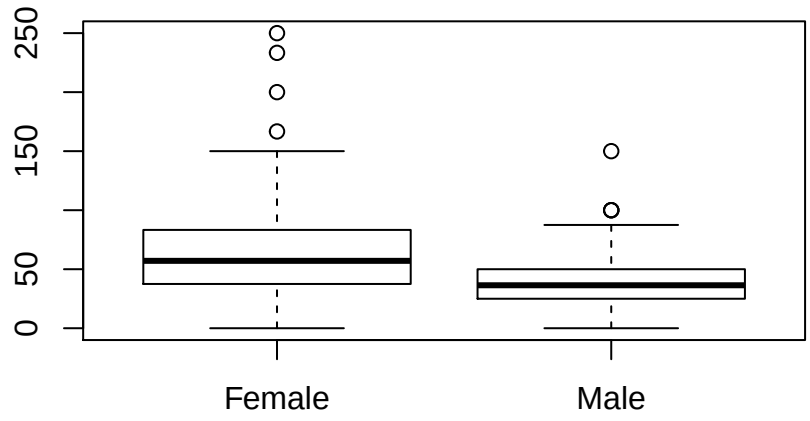
In the Functional polymorphisms Associated with Human Muscle Size and Strength study (FAMuSS), researchers investigated the relationship between muscle strength and genotype at a location on the ACTN3 gene. The `famuss` dataset in `oibiostat` contains a subset of the data collected from the study, which includes additional demographic and phenotypic information. The percent change in non-dominant arm strength, comparing strength after training to before training, is stored as `ndrm.ch`.

6. Load and view the data. Create a plot that shows the association between `ndrm.ch` and `sex`. Describe what you see.

   The median change in non-dominant arm strength is higher for females than for males. There is also greater variance in `ndrm.ch` for females than for males.

```
#load the data
library(oibiostat)
data("famuss")

#create a plot
boxplot(famuss$ndrm.ch ~ famuss$sex)
```




7. Conduct a hypothesis test to address the question of interest. Let $\alpha = 0.05$.

   a) Suppose the parameter $\mu_F$ is the population mean change in non-dominant arm strength for women, and $\mu_M$ is the population mean change in non-dominant arm strength for men. State the null and alternative hypotheses.

   The null hypothesis is that the population mean change in arm strength does not differ between men and women, $H_0 : \mu_F = \mu_M$. The alternative hypothesis is that the population

mean change in arm strength does differ between men and women, $H_A : \mu_F \neq \mu_M$.

b) Calculate the test statistic for the difference in means of the two groups, where $\bar{x}_1$ and $\bar{x}_2$ are the means of each sample, $s_1$ and $s_2$ are the standard deviations of each sample, and $n_1$ and $n_2$ are the number of observations per sample.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(62.93 - 39.24) - (0)}{\sqrt{\frac{36.52^2}{353} + \frac{20.60^2}{242}}} = 10.07$$

```
#define categories for sorting ndrm.ch
female = (famuss$sex == "Female")
male = (famuss$sex == "Male")

#set parameters
x.bar.1 = mean(famuss$ndrm.ch[female])
s.1 = sd(famuss$ndrm.ch[female])
n.1 = length((famuss$ndrm.ch[female]))

x.bar.2 = mean(famuss$ndrm.ch[male])
s.2 = sd(famuss$ndrm.ch[male])
n.2 = length((famuss$ndrm.ch[male]))

#calculate t-statistic
t.num = (x.bar.1 - x.bar.2)
t.den = sqrt( ((s.1^2)/n.1) + ((s.2^2)/n.2) )
t.stat = t.num/t.den
t.stat
```

```
## [1] 10.07294
```

c) Calculate the $p$-value. The degrees of freedom for the $t$-statistic in this setting can be approximated as the smaller of $n_1 - 1$ and $n_2 - 1$; i.e., $\min(n_1 - 1, n_2 - 1)$.

The $p$-value is $P(T \geq |10.07|) = 2 \times P(T \geq 10.07) = 3.82 \times 10^{-20}$.

```
#calculate the p-value
2*pt(t.stat, df = min(n.1 - 1, n.2 - 1), lower.tail = FALSE)
```

```
## [1] 3.823129e-20
```

d) Draw a conclusion.

The $p$-value is much smaller than $\alpha$; the results are significant at the 0.05 significance level and there is evidence to reject the null hypothesis. The data suggest that the population mean change in non-dominant arm strength for females is greater than that for males.

8. Calculate a 95% confidence interval. Interpret the interval in the context of the data.

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

$$(62.93 - 39.24) \pm 1.97 \times \sqrt{\frac{36.52^2}{353} + \frac{20.60^2}{242}}$$

$$(19.06, 28.33) \ \%$$

We are 95% confident that the true difference in mean change in non-dominant arm strength between women and men lies within the interval (19.06, 28.33) %.

```
#use r as a calculator
x.bar.1 = mean(famuss$ndrm.ch[female])
s.1 = sd(famuss$ndrm.ch[female])
n.1 = length((famuss$ndrm.ch[female]))

x.bar.2 = mean(famuss$ndrm.ch[male])
s.2 = sd(famuss$ndrm.ch[male])
n.2 = length((famuss$ndrm.ch[male]))

t.star = qt(0.975, df = min(n.1 - 1, n.2 - 2))
m = t.star * sqrt( ((s.1^2)/n.1) + ((s.2^2)/n.2) )

x.bar.1 - x.bar.2 - m; x.bar.1 - x.bar.2 + m
```

```
## [1] 19.05877
```

```
## [1] 28.32537
```

9. Verify the answers to Questions 7 and 8 using `t.test()`. Where do the values returned by `t.test()` differ from the hand calculation?

   While the $t$-statistic is the same, the degrees of freedom differ. The $p$-value in both cases is vanishingly small, and `t.test()` can only report that the $p$-value is smaller than $2.2 \times 10^{-16}$. The values for the confidence interval are very close, and only differ by the hundredths decimal place.

   R calculates the degrees of freedom according to the equation below, which is a better approximation than using the minimum of $n_1 - 1$ and $n_2 - 1$.

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

```
#option 1: use categories as defined earlier
t.test(famuss$ndrm.ch[female], famuss$ndrm.ch[male], paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  famuss$ndrm.ch[female] and famuss$ndrm.ch[male]
## t = 10.073, df = 574.01, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19.07240 28.31175
## sample estimates:
## mean of x mean of y
##  62.92720  39.23512
```

```
#option 2: tilde syntax
t.test(famuss$ndrm.ch ~ famuss$sex, alternative = "two.sided", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  famuss$ndrm.ch by famuss$sex
## t = 10.073, df = 574.01, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  19.07240 28.31175
## sample estimates:
## mean in group Female   mean in group Male
##             62.92720             39.23512
```

```
#calculating the exact degrees of freedom
x.bar.1 = mean(famuss$ndrm.ch[female])
s.1 = sd(famuss$ndrm.ch[female])
n.1 = length((famuss$ndrm.ch[female]))

x.bar.2 = mean(famuss$ndrm.ch[male])
s.2 = sd(famuss$ndrm.ch[male])
n.2 = length((famuss$ndrm.ch[male]))

df.num = ( ((s.1^2)/n.1) +((s.2)^2)/n.2  )^2
df.den = ((s.1^2 / n.1)^2)/(n.1 - 1) + ((s.2^2 / n.2)^2)/(n.2 - 1)
df = df.num/df.den
df
```

```
## [1] 574.0122
```

**Additional practice**

10. Suppose the swim data had been incorrectly identified as independent two-group data.

    a) Use `t.test()` to analyze the data accordingly and summarize the results. Do the results suggest a different conclusion than those of the original analysis?

    If the swim data are analyzed as if they are independent two-group data, then the $p$-value is 0.31, which is greater than 0.05. The results would not have indicated the data were extreme enough to reject the null hypothesis of no difference in swim velocities.

```
t.test(swim$wet.suit.velocity, swim$swim.suit.velocity,
   mu = 0, alternative = "two.sided", paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  swim$wet.suit.velocity and swim$swim.suit.velocity
## t = 1.3688, df = 21.974, p-value = 0.1849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03992937  0.19492937
## sample estimates:
## mean of x mean of y
##  1.506667  1.429167
```

    b) Conjecture as to why a paired study design has more capacity to show evidence of a difference between two groups than an independent group design. (Hint: think about the variance between individuals, versus the variance between two measurements on the same individual).

    Consider an independent group design for the swim study in which one half of the swimmers complete a 1500m swim wearing a wetsuit and the other half swim wearing a swimsuit. In order for there to be sufficient evidence of a difference, the difference in means must be large relative to the variance in swim velocities within groups (as measured by the standard error of the test statistic).

    In a paired study design, each swimmer acts as their own control. Rather than measure the effect of wearing a swim suit by comparing the average difference between two groups, the pairing allows for a difference to be calculated at the individual level, for each participant. A paired design leverages the fact that two measurements on one individual will tend to be correlated, and more similar than two measurements made on separate individuals.

    Note: The concept of a statistical test having more capacity to show evidence of a difference (when the alternative hypothesis is true and a difference exists at the population level) is referred to as power. Power will be explored in a later section of the chapter (and lab exercise).

11. Chapter 1 (Section 7.1, Lab 2) presented an exploratory analysis of the relationship between age, ethnicity, and the amount of expenditures for supporting developmentally disabled residents in the State of California. When age is ignored, the expenditures per consumer is larger on average for White non-Hispanics than Hispanics, but the average differences by ethnicity were much smaller within age cohorts. Hypothesis testing can be used to conduct

a more formal analysis of the differences in expenditures by ethnicity, both overall (i.e., ignoring age) and within age cohorts.

The data are in the `dds.discr` dataset in `oibiostat`. Descriptions of the variables are provided in the documentation file.

a) Is there evidence of a difference in mean expenditures by ethnic group, comparing Hispanics to White non-Hispanics? There is substantial skew in the distribution of expenditures within each group, so it is advisable to apply a natural log transformation before conducting a $t$-test. Summarize the results of the test.

When ignoring age, there is significant evidence of a difference in mean expenditures between Hispanics and White non-Hispanics. It appears that on average, White non-Hispanics receive a higher amount of developmental disability support from the state of California ($\overline{x} > \overline{y}$).

```
#load the data
library(oibiostat)
data("dds.discr")

#apply transformation
dds.discr$log.expenditures = log(dds.discr$expenditures)

#conduct test
t.test(dds.discr$log.expenditures[dds.discr$ethnicity == "White not Hispanic"],
   dds.discr$log.expenditures[dds.discr$ethnicity == "Hispanic"])
```

```
##
##  Welch Two Sample t-test
##
## data:  dds.discr$log.expenditures[dds.discr$ethnicity == "White not Hispanic"] and dds.discr$
## t = 10.107, df = 770.48, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7357922 1.0905107
## sample estimates:
## mean of x mean of y
##  9.471642  8.558491
```

b) One way to account for the effect of age is to compare mean expenditures within age cohorts. When comparing individuals of similar ages, are the differences in mean expenditures by ethnic group larger than would be expected by chance alone?

   Conduct two $t$-tests to examine the evidence against the null hypothesis of no difference in mean expenditures within the two largest age cohorts: 13-17 years and 22-50 years. Summarize the results.

The $p$-values are large (0.75 and 0.51); it is not unlikely to see the observed discrepancy in means if the population means are actually the same. There is not evidence of a difference between mean expenditures in Hispanics and White non-Hispanics within either age cohort. This inference-based approach supports the conclusions from the earlier exploratory lab.

```r
#create subsets
ages.13to17 = dds.discr[dds.discr$age.cohort == "13-17", ]
ages.22to50 = dds.discr[dds.discr$age.cohort == "22-50", ]

#conduct tests
t.test(ages.13to17$expenditures[ages.13to17$ethnicity == "White not Hispanic"],
   ages.13to17$expenditures[ages.13to17$ethnicity == "Hispanic"])
```

```
##
##  Welch Two Sample t-test
##
## data:  ages.13to17$expenditures[ages.13to17$ethnicity == "White not Hispanic"] and ages.13to1
## t = -0.31779, df = 127.8, p-value = 0.7512
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -367.9960  266.1493
## sample estimates:
## mean of x mean of y
##  3904.358  3955.282
```

```r
t.test(ages.22to50$expenditures[ages.22to50$ethnicity == "White not Hispanic"],
   ages.22to50$expenditures[ages.22to50$ethnicity == "Hispanic"])
```

```
##
##  Welch Two Sample t-test
##
## data:  ages.22to50$expenditures[ages.22to50$ethnicity == "White not Hispanic"] and ages.22to5
## t = -0.65855, df = 67.687, p-value = 0.5124
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2968.306  1495.321
## sample estimates:
## mean of x mean of y
##  40187.62  40924.12
```