Statistical inference: one and two-sample t-tests

Statistical Inference and Science

- Previously: descriptive statistics. "Here are data; what do they say?".
- May need to take some action based on information in data.
- Or want to generalize beyond data (sample) to larger world (population).
- Science: first guess about how world works.
- Then collect data, by sampling.
- Is guess correct (based on data) for whole world, or not?

Sample data are imperfect

- Sample data never entirely represent what you're observing.
- There is always random error present.
- Thus you can never be entirely certain about your conclusions.
- The Toronto Blue Jays' average home attendance in part of 2015 season was 25,070 (up to May 27 2015, from baseball-reference.com).
- Does that mean the attendance at every game was exactly 25,070? Certainly not. Actual attendance depends on many things, eg.:
  - how well the Jays are playing
  - the opposition
  - day of week
  - weather
  - random chance

Packages for this section

```
library(tidyverse)
```

Reading the attendances

…as a `.csv` file:

```
my_url <- "http://ritsokiguess.site/datafiles/jays15-home.csv"
jays <- read_csv(my_url)
jays
```

```
# A tibble: 25 x 21
     row  game date   box   team  venue opp   result  runs Oppruns innings wl   
   <dbl> <dbl> <chr>  <chr> <chr> <lgl> <chr> <chr>  <dbl>   <dbl>   <dbl> <chr>
 1    82     7 Monda~ boxs~ TOR   NA    TBR   L          1       2      NA 4-3  
 2    83     8 Tuesd~ boxs~ TOR   NA    TBR   L          2       3      NA 4-4  
 3    84     9 Wedne~ boxs~ TOR   NA    TBR   W         12       7      NA 5-4  
 4    85    10 Thurs~ boxs~ TOR   NA    TBR   L          2       4      NA 5-5  
 5    86    11 Frida~ boxs~ TOR   NA    ATL   L          7       8      NA 5-6  
 6    87    12 Satur~ boxs~ TOR   NA    ATL   W-wo       6       5      10 6-6  
 7    88    13 Sunda~ boxs~ TOR   NA    ATL   L          2       5      NA 6-7  
 8    89    14 Tuesd~ boxs~ TOR   NA    BAL   W         13       6      NA 7-7  
 9    90    15 Wedne~ boxs~ TOR   NA    BAL   W          4       2      NA 8-7  
10    91    16 Thurs~ boxs~ TOR   NA    BAL   W          7       6      NA 9-7  
# i 15 more rows
# i 9 more variables: position <dbl>, gb <chr>, winner <chr>, loser <chr>,
#   save <chr>, `game time` <time>, Daynight <chr>, attendance <dbl>,
#   streak <chr>
```

Another way

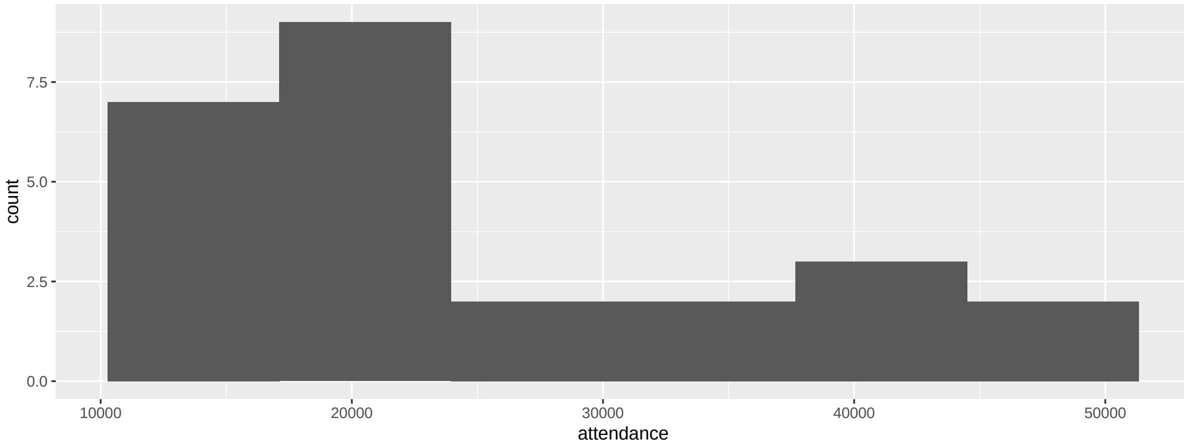
- This is a big data set: only 25 observations, but a lot of *variables*.
- To see the first few values in all the variables, can also use `glimpse`:

```
glimpse(jays)
```

```
Rows: 25
Columns: 21
$ row         <dbl> 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94~
$ game        <dbl> 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 27, 28, 29, 3~
$ date        <chr> "Monday, Apr 13", "Tuesday, Apr 14", "Wednesday, A~
$ box         <chr> "boxscore", "boxscore", "boxscore", "boxscore", "b~
$ team        <chr> "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "~
$ venue       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ opp         <chr> "TBR", "TBR", "TBR", "TBR", "ATL", "ATL", "ATL", "~
$ result      <chr> "L", "L", "W", "L", "L", "W-wo", "L", "W", "W", "W~
$ runs        <dbl> 1, 2, 12, 2, 7, 6, 2, 13, 4, 7, 3, 3, 5, 7, 7, 3, ~
$ Oppruns     <dbl> 2, 3, 7, 4, 8, 5, 5, 6, 2, 6, 1, 6, 1, 0, 1, 6, 6,~
$ innings     <dbl> NA, NA, NA, NA, NA, 10, NA, NA, NA, NA, NA, NA, NA~
$ wl          <chr> "4-3", "4-4", "5-4", "5-5", "5-6", "6-6", "6-7", "~
$ position    <dbl> 2, 3, 2, 4, 4, 3, 4, 2, 2, 1, 4, 5, 3, 3, 3, 3, 5,~
$ gb          <chr> "1", "2", "1", "1.5", "2.5", "1.5", "1.5", "2", "1~
$ winner      <chr> "Odorizzi", "Geltz", "Buehrle", "Archer", "Martin"~
$ loser       <chr> "Dickey", "Castro", "Ramirez", "Sanchez", "Cecil",~
$ save        <chr> "Boxberger", "Jepsen", NA, "Boxberger", "Grilli", ~
$ `game time` <time> 02:30:00, 03:06:00, 03:02:00, 03:00:00, 03:09:00,~
$ Daynight    <chr> "N", "N", "N", "N", "N", "D", "D", "N", "N", "N", ~
$ attendance  <dbl> 48414, 17264, 15086, 14433, 21397, 34743, 44794, 1~
$ streak      <chr> "-", "--", "+", "-", "--", "+", "-", "+", "++", "+~
```

Attendance histogram

```
ggplot(jays, aes(x = attendance)) + geom_histogram(bins = 6)
```



Comments

- Attendances have substantial variability, ranging from just over 10,000 to around 50,000.
- Distribution somewhat skewed to right (but no outliers).
- These are a sample of "all possible games" (or maybe "all possible games played in April and May"). What can we say about mean attendance in all possible games based on this evidence?
- Think about:
  - Confidence interval
  - Hypothesis test.

Getting CI for mean attendance

- `t.test` function does CI and test. Look at CI first:

```
t.test(jays$attendance)
```

```

	One Sample t-test

data:  jays$attendance
t = 11.389, df = 24, p-value = 3.661e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 20526.82 29613.50
sample estimates:
mean of x 
 25070.16 
```

- From 20,500 to 29,600.

Or, 90% CI

- by including a value for conf.level:

```
t.test(jays$attendance, conf.level = 0.90)
```

```
    One Sample t-test

data:  jays$attendance
t = 11.389, df = 24, p-value = 3.661e-11
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 21303.93 28836.39
sample estimates:
mean of x
 25070.16
```

- From 21,300 to 28,800. (Shorter, as it should be.)

Comments

- Need to say "column attendance within data frame `jays`" using $.
- 95% CI from about 20,000 to about 30,000.
- Not estimating mean attendance well at all!
- Generally want confidence interval to be shorter, which happens if:
  - SD smaller
  - sample size bigger
  - confidence level smaller
- Last one is a cheat, really, since reducing confidence level increases chance that interval won't contain pop. mean at all!

Another way to access data frame columns

```
with(jays, t.test(attendance))
```

```
    One Sample t-test

data:  attendance
t = 11.389, df = 24, p-value = 3.661e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 20526.82 29613.50
sample estimates:
```

```
mean of x
 25070.16
```

Hypothesis test

- CI answers question "what is the mean?"
- Might have a value μ in mind for the mean, and question "Is the mean equal to μ, or not?"
- For example, 2014 average attendance was 29,327.
- "Is the mean this?" answered by **hypothesis test**.
- Value being assessed goes in **null hypothesis**: here, $H_0 : \mu = 29327$.
- **Alternative hypothesis** says how null might be wrong, eg. $H_a : \mu \neq 29327$.
- Assess evidence against null. If that evidence strong enough, *reject null hypothesis;* if not, *fail to reject null hypothesis* (sometimes *retain null*).
- Note asymmetry between null and alternative, and utter absence of word "accept".

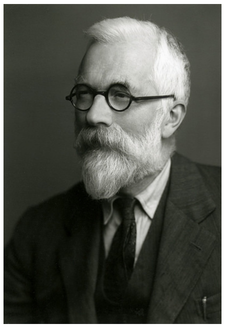
α and errors

- Hypothesis test ends with decision:
  - reject null hypothesis
  - do not reject null hypothesis.
- but decision may be wrong:

	Decision	
Truth	Do not reject	Reject null
Null true	Correct	Type I error
Null false	Type II error	Correct

- Either type of error is bad, but for now focus on controlling Type I error: write $\alpha =$ P(type I error), and devise test so that α small, typically 0.05.
- That is, **if null hypothesis true**, have only small chance to reject it (which would be a mistake).
- Worry about type II errors later (when we consider power of test).

Why 0.05? This man.

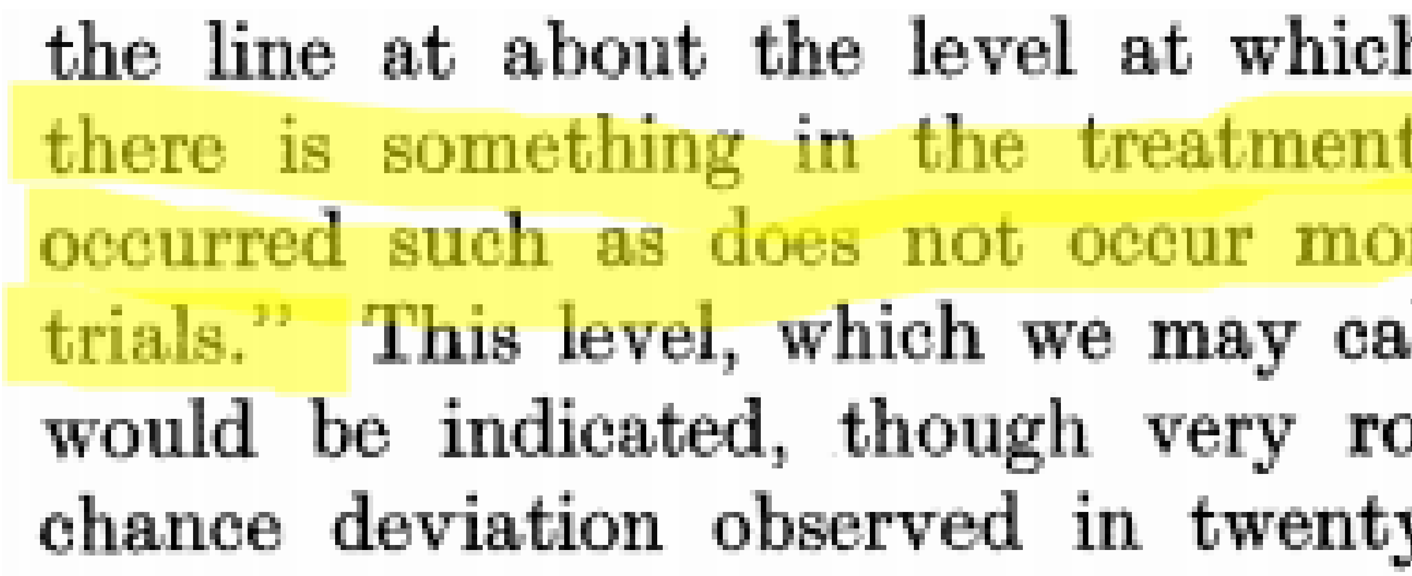


- analysis of variance
- Fisher information
- Linear discriminant analysis
- Fisher's z-transformation
- Fisher-Yates shuffle
- Behrens-Fisher problem

Sir Ronald A. Fisher, 1890–1962.

Why 0.05? (2)

- From The Arrangement of Field Experiments (1926):



the line at about the level at whic
there is something in the treatment
occurred such as does not occur mo
trials." This level, which we may ca
would be indicated, though very ro
chance deviation observed in twenty

- and

If one in twenty does not seem hig
if we prefer it, draw the line at one
point), or one in a hundred (the 1 per
the writer prefers to set a low standa
5 per cent. point, and ignore entirely
reach this level. A scientific fact shou
mentally established only if a prope
rarely fails to give this level of signi

Three steps:

- from data to test statistic
  - how far are data from null hypothesis
- from test statistic to P-value
  - how likely are you to see "data like this" **if the null hypothesis is true**
- from P-value to decision
  - reject null hypothesis if P-value small enough, fail to reject it otherwise

Using `t.test`:

```
t.test(jays$attendance, mu=29327)
```

```
	One Sample t-test

data:  jays$attendance
t = -1.9338, df = 24, p-value = 0.06502
alternative hypothesis: true mean is not equal to 29327
95 percent confidence interval:
 20526.82 29613.50
sample estimates:
mean of x
 25070.16
```

- See test statistic -1.93, P-value 0.065.
- Do not reject null at $\alpha = 0.05$: no evidence that mean attendance has changed.

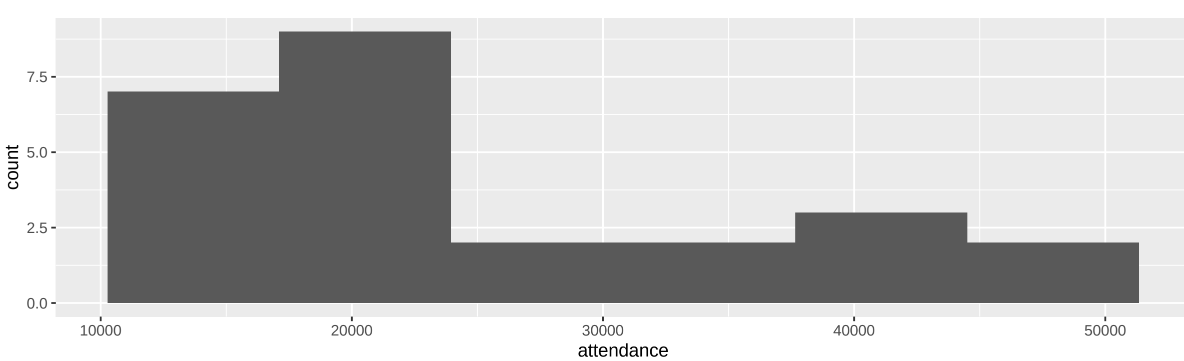
Assumptions

- Theory for t-test: assumes normally-distributed data.
- What actually matters is sampling distribution of sample mean: if this is approximately normal, t-test is OK, even if data distribution is not normal.
- Central limit theorem: if sample size large, sampling distribution approx. normal even if data distribution somewhat non-normal.
- So look at shape of data distribution, and make a call about whether it is normal enough, given the sample size.

Blue Jays attendances again:

- You might say that this is not normal enough for a sample size of $n = 25$, in which case you don't trust the t-test result:

```
ggplot(jays, aes(x = attendance)) + geom_histogram(bins = 6)
```



Another example: learning to read

- You devised new method for teaching children to read.
- Guess it will be more effective than current methods.
- To support this guess, collect data.
- Want to generalize to "all children in Canada".
- So take random sample of all children in Canada.
- Or, argue that sample you actually have is "typical" of all children in Canada.
- Randomization (1): whether or not a child in sample or not has nothing to do with anything else about that child.
- Randomization (2): randomly choose whether each child gets new reading method (t) or standard one (c).

Reading in data

- File at http://ritsokiguess.site/datafiles/drp.txt.
- Proper reading-in function is `read_delim` (check file to see)
- Read in thus:

```
my_url <- "http://ritsokiguess.site/datafiles/drp.txt"
kids <- read_delim(my_url," ")
```

The data

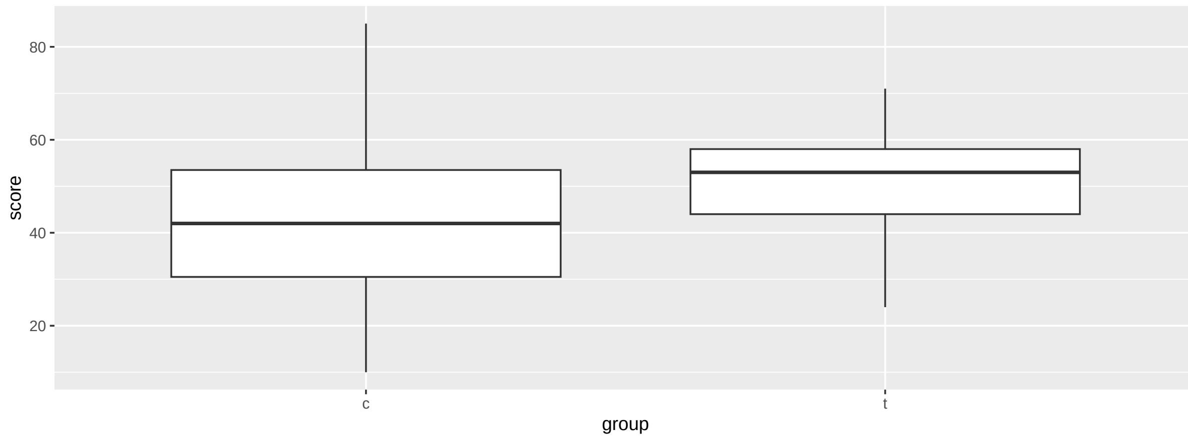
```
kids
```

```
# A tibble: 44 x 2
   group score
   <chr> <dbl>
 1 t        24
 2 t        61
 3 t        59
 4 t        46
 5 t        43
 6 t        44
 7 t        52
 8 t        43
 9 t        58
10 t        67
# i 34 more rows
```

In `group`, `t` is "treatment" (the new reading method) and `c` is "control" (the old one).

Boxplots

```
ggplot(kids, aes(x = group, y = score)) + geom_boxplot()
```



Two kinds of two-sample t-test

- Do the two groups have same spread (SD, variance)?
  - If yes (shaky assumption here), can use pooled t-test.
  - If not, use Welch-Satterthwaite t-test (safe).
- Pooled test derived in STAB57 (easier to derive).
- Welch-Satterthwaite is test used in STAB22 and is generally safe.
- Assess (approx) equality of spreads using boxplot.

The (Welch-Satterthwaite) t-test

- `c` (control) before `t` (treatment) alphabetically, so proper alternative is "less".
- R does Welch-Satterthwaite test by default
- Answer to "does the new reading program really help?"
- (in a moment) how to get R to do pooled test?

Welch-Satterthwaite

```
t.test(score ~ group, data = kids, alternative = "less")
```

```

    Welch Two Sample t-test

data:  score by group
t = -2.3109, df = 37.855, p-value = 0.01319
alternative hypothesis: true difference in means between group c and group t is less than 0
95 percent confidence interval:
      -Inf -2.691293
sample estimates:
mean in group c mean in group t 
       41.52174        51.47619 
```

The pooled t-test

```
t.test(score ~ group, data = kids,
       alternative = "less", var.equal = TRUE)
```

```

    Two Sample t-test

data:  score by group
t = -2.2666, df = 42, p-value = 0.01431
alternative hypothesis: true difference in means between group c and group t is less than 0
95 percent confidence interval:
      -Inf -2.567497
sample estimates:
mean in group c mean in group t 
       41.52174        51.47619 
```

Two-sided test; CI

- To do 2-sided test, leave out `alternative`:

```
t.test(score ~ group, data = kids)
```

```
    Welch Two Sample t-test

data:  score by group
t = -2.3109, df = 37.855, p-value = 0.02638
alternative hypothesis: true difference in means between group c and group t is not equal to
95 percent confidence interval:
 -18.67588  -1.23302
sample estimates:
mean in group c mean in group t
       41.52174        51.47619
```

Comments:

- P-values for pooled and Welch-Satterthwaite tests very similar (even though the pooled test seemed inferior): 0.013 vs. 0.014.
- Two-sided test also gives CI: new reading program increases average scores by somewhere between about 1 and 19 points.
- Confidence intervals inherently two-sided, so do 2-sided test to get them.

Jargon for testing

- Alternative hypothesis: what we are trying to prove (new reading program is effective).
- Null hypothesis: "there is no difference" (new reading program no better than current program). Must contain "equals".
- One-sided alternative: trying to prove better (as with reading program).
- Two-sided alternative: trying to prove different.
- Test statistic: something expressing difference between data and null (eg. difference in sample means, t statistic).
- P-value: probability of observing test statistic value as extreme or more extreme, if null is true.
- Decision: either reject null hypothesis or do not reject null hypothesis. **Never "accept"**.

Logic of testing

- Work out what would happen if null hypothesis were true.
- Compare to what actually did happen.
- If these are too far apart, conclude that null hypothesis is not true after all. (Be guided by P-value.)
- As applied to our reading programs:

  - If reading programs equally good, expect to see a difference in means close to 0.
  - Mean reading score was 10 higher for new program.
  - Difference of 10 was unusually big (P-value small from t-test). So conclude that new reading program is effective.

- Nothing here about what happens if null hypothesis is false. This is power and type II error probability.