Statistical Inference: one- and two-sample inference

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

```r
library(tidyverse)
# library(smmr)
# library(PMCMRplus)
```

Reading the attendances

...as a `.csv` file:

```
jays <- read_csv("jays15-home.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   row = col_double(),
##   game = col_double(),
##   venue = col_logical(),
##   runs = col_double(),
##   Oppruns = col_double(),
##   innings = col_double(),
##   position = col_double(),
##   `game time` = col_time(format = ""),
##   attendance = col_double()
## )
```

Taking a look (tiny)

```
jays
```

row	game	date	box	team	venue	opp	result	runs	Oppruns	innings	wl	position	gb	winner	loser	save	g t
82	7	Monday, Apr 13	boxscore	TOR	NA	TBR	L	1	2	NA	4-3	2	1	Odorizzi	Dickey	Boxberge	0
83	8	Tuesday, Apr 14	boxscore	TOR	NA	TBR	L	2	3	NA	4-4	3	2	Geltz	Castro	Jepsen	0
84	9	Wednesday, Apr 15	boxscore	TOR	NA	TBR	W	12	7	NA	5-4	2	1	Buehrle	Ramirez	NA	0
85	10	Thursday, Apr 16	boxscore	TOR	NA	TBR	L	2	4	NA	5-5	4	1.5	Archer	Sanchez	Boxberge	0
86	11	Friday, Apr 17	boxscore	TOR	NA	ATL	L	7	8	NA	5-6	4	2.5	Martin	Cecil	Grilli	0
87	12	Saturday, Apr 18	boxscore	TOR	NA	ATL	W-wo	6	5	10	6-6	3	1.5	Cecil	Marimon	NA	0
88	13	Sunday, Apr 19	boxscore	TOR	NA	ATL	L	2	5	NA	6-7	4	1.5	Miller	Norris	Grilli	0
89	14	Tuesday, Apr 21	boxscore	TOR	NA	BAL	W	13	6	NA	7-7	2	2	Buehrle	Norris	NA	0
90	15	Wednesday, Apr 22	boxscore	TOR	NA	BAL	W	4	2	NA	8-7	2	1	Sanchez	Jimenez	Castro	0
91	16	Thurs-	boxs-	TOR	NA	BAL	W	7	6	NA	9-	1	Tied	Hutchi-	Till-	Cas-	0

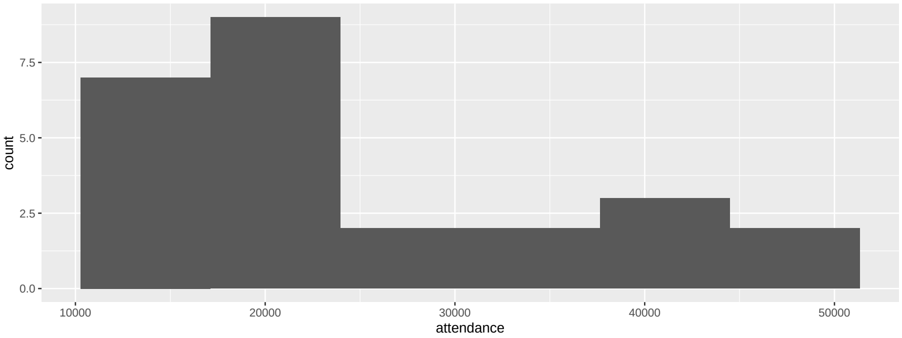
Another way

```
glimpse(jays)
```

```
## Rows: 25
## Columns: 21
## $ row         <dbl> 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100...
## $ game        <dbl> 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 27, 28, 29, 30, 31, 32, 40, 41, 42, 43...
## $ date        <chr> "Monday, Apr 13", "Tuesday, Apr 14", "Wednesday, Apr 15", "Thursday, Apr 16...
## $ box         <chr> "boxscore", "boxscore", "boxscore", "boxscore", "boxscore", "boxscore", "bo...
## $ team        <chr> "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "TOR", "TOR"...
## $ venue       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ opp         <chr> "TBR", "TBR", "TBR", "TBR", "ATL", "ATL", "ATL", "BAL", "BAL", "BAL", "NYY"...
## $ result      <chr> "L", "L", "W", "L", "L", "W-wo", "L", "W", "W", "W", "W", "L", "W", "W", "W...
## $ runs        <dbl> 1, 2, 12, 2, 7, 6, 2, 13, 4, 7, 3, 3, 5, 7, 7, 3, 10, 2, 3, 8, 3, 2, 8, 6, 10
## $ Oppruns     <dbl> 2, 3, 7, 4, 8, 5, 5, 6, 2, 6, 1, 6, 1, 0, 1, 6, 6, 3, 4, 4, 4, 3, 2, 0, 9
## $ innings     <dbl> NA, NA, NA, NA, NA, 10, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ wl          <chr> "4-3", "4-4", "5-4", "5-5", "5-6", "6-6", "6-7", "7-7", "8-7", "9-7", "13-1...
## $ position    <dbl> 2, 3, 2, 4, 4, 3, 4, 2, 2, 1, 4, 5, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 5, 4
## $ gb          <chr> "1", "2", "1", "1.5", "2.5", "1.5", "1.5", "2", "1", "Tied", "3.5", "4.5", ...
## $ winner      <chr> "Odorizzi", "Geltz", "Buehrle", "Archer", "Martin", "Cecil", "Miller", "Bue...
## $ loser       <chr> "Dickey", "Castro", "Ramirez", "Sanchez", "Cecil", "Marimon", "Norris", "No...
## $ save        <chr> "Boxberger", "Jepsen", NA, "Boxberger", "Grilli", NA, "Grilli", NA, "Castro...
## $ `game time` <time> 02:30:00, 03:06:00, 03:02:00, 03:00:00, 03:09:00, 02:41:00, 02:41:00, 02:5...
## $ Daynight    <chr> "N", "N", "N", "N", "N", "D", "D", "N", "N", "N", "N", "N", "N", "N", "D", ...
## $ attendance  <dbl> 48414, 17264, 15086, 14433, 21397, 34743, 44794, 14184, 15606, 18581, 19217...
## $ 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

```r
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".

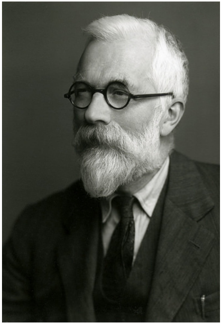
α and errors

- 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.



Responsible for:

- 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 which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." This level, which we may call the 5 per cent. point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials. To

- and

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. The very high

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.

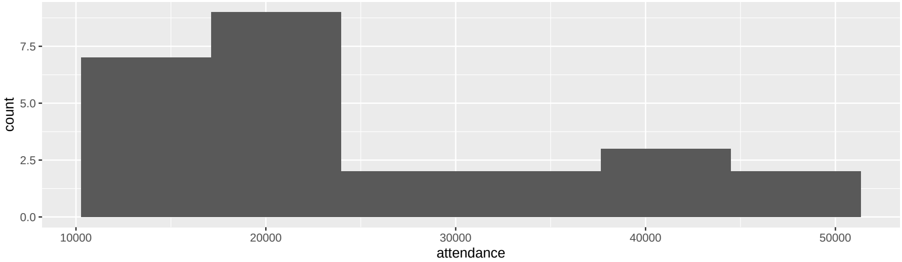
Assumptions

- 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://www.utsc.utoronto.ca/~butler/c32/drp.txt.
- Proper reading-in function is `read_delim` (check file to see)
- Read in thus:

```
my_url <- "http://www.utsc.utoronto.ca/~butler/c32/drp.txt"
kids <- read_delim(my_url," ")
```

```
## Parsed with column specification:
## cols(
##   group = col_character(),
##   score = col_double()
## )
```

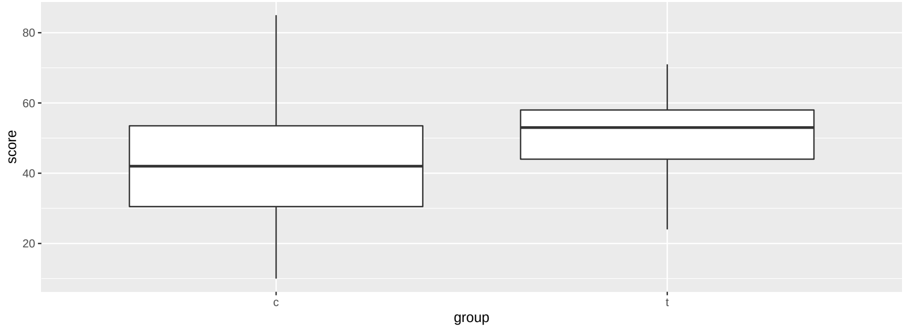
The data (some)

```
kids
```

group	score
t	24
t	61
t	59
t	46
t	43
t	44
t	52
t	43
t	58
t	67
t	62
t	57
t	71

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

```r
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 is less th
## 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

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

```
##
##  Two Sample t-test
##
## data:  score by group
## t = -2.2666, df = 42, p-value = 0.01431
## alternative hypothesis: true difference in means is less th
## 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 is not equ
## 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.