Stat 140: Inference for a Difference in Means

Evan Ray

November 27, 2017

Jumping Frog Jubilee

Introduction

Quote from Astley et al., 2013. Chasing maximal performance: a cautionary tale from the celebrated jumping frogs of Calaveras County. The Journal of Experimental Biology 216, 3947-3953.

> We recorded video of 3124 bullfrog jumps over the course of the 4-day contest at the Calaveras County Jumping Frog Jubilee, and determined jump distance from these images and a calibration of the jump arena. Frogs were divided into two groups: 'rental' frogs collected by fair organizers and jumped by the general public, and frogs collected and jumped by experienced, 'professional' teams.

Read in the data, and subset to non-professional frog-jumpers:

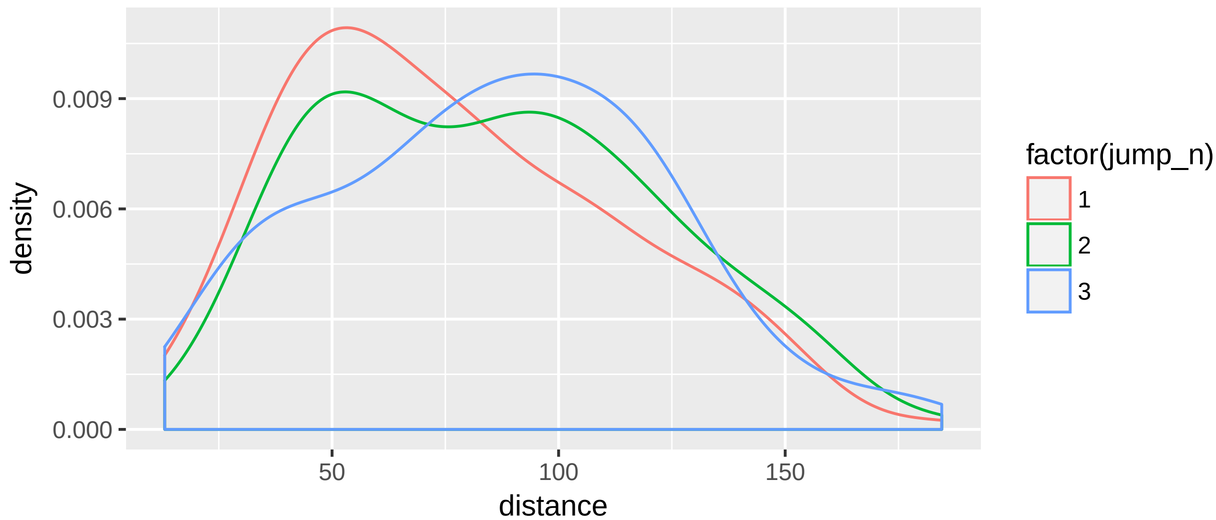
```
frogs <- read.csv("https://mhc-stat140-2017.github.io/data/misc/frogs/frogs.csv")
head(frogs)
```

```
##   id jump_n frog_type distance duration distance_3 distance_3_off
## 1  1      1       pro  165.950  0.58333          0             -1
## 2  1      2       pro  177.480  0.71667          0             -1
## 3  1      3       pro    0.000  0.00000          0             -1
## 4  2      1       pro   27.158  0.43333          0             -1
## 5  2      2       pro    0.000  0.00000          0             -1
## 6  2      3       pro    0.000  0.00000          0             -1
##   distance_rel day angle_01 angle_10 angle_00 velocity_01 velocity_10
## 1       1.0000   1 28.85564 24.90057 33.03045    3.711031    3.876228
## 2       1.0695   1 41.44158 37.19646 45.62517    3.680306    3.700908
## 3       0.0000   1       NA       NA       NA          NA          NA
## 4       1.0000   1       NA       NA       NA          NA          NA
## 5       0.0000   1       NA       NA       NA          NA          NA
## 6       0.0000   1       NA       NA       NA          NA          NA
##   velocity_00
## 1    3.599155
## 2    3.702692
## 3          NA
## 4          NA
## 5          NA
## 6          NA
```

```
frogs_amateur <- frogs %>%
  filter(frog_type %in% c("rental", "individual") & jump_n <= 3 & distance > 0)
```

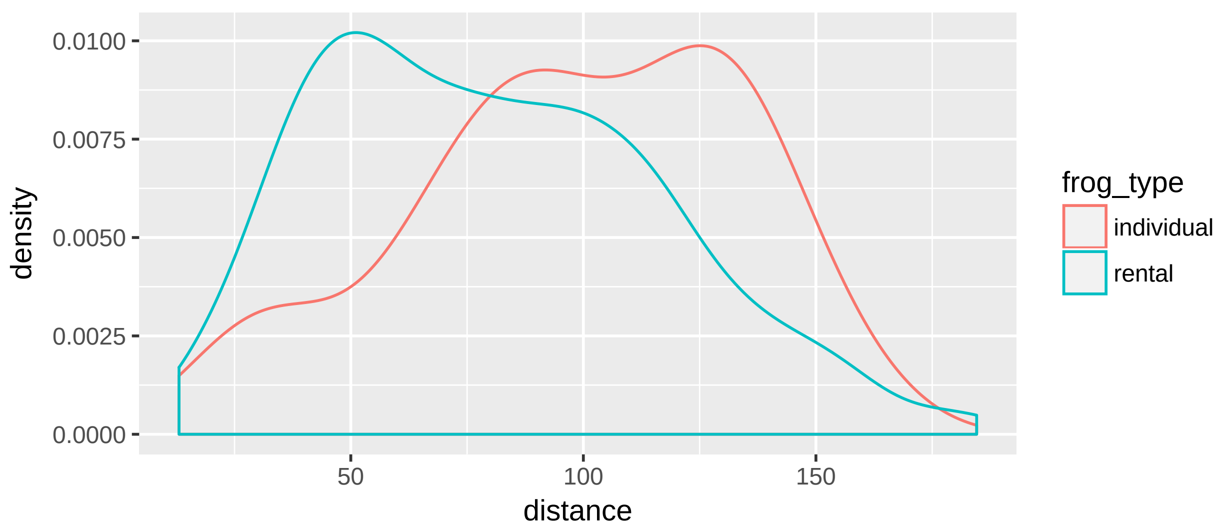
Plot jump distance by jump number:

```
ggplot(frogs_amateur, aes(x = distance, color = factor(jump_n))) +
  geom_density()
```



Plot jump distance by frog type:

```
ggplot(frogs_amateur, aes(x = distance, color = frog_type)) +
  geom_density()
```



Example 1: Are mean frog jump distances the same for "rental" frogs and for frogs brought to the fair by amateur "individual"s?

State Null and Alternative Hypotheses

Check Assumptions for Two-Sample t test

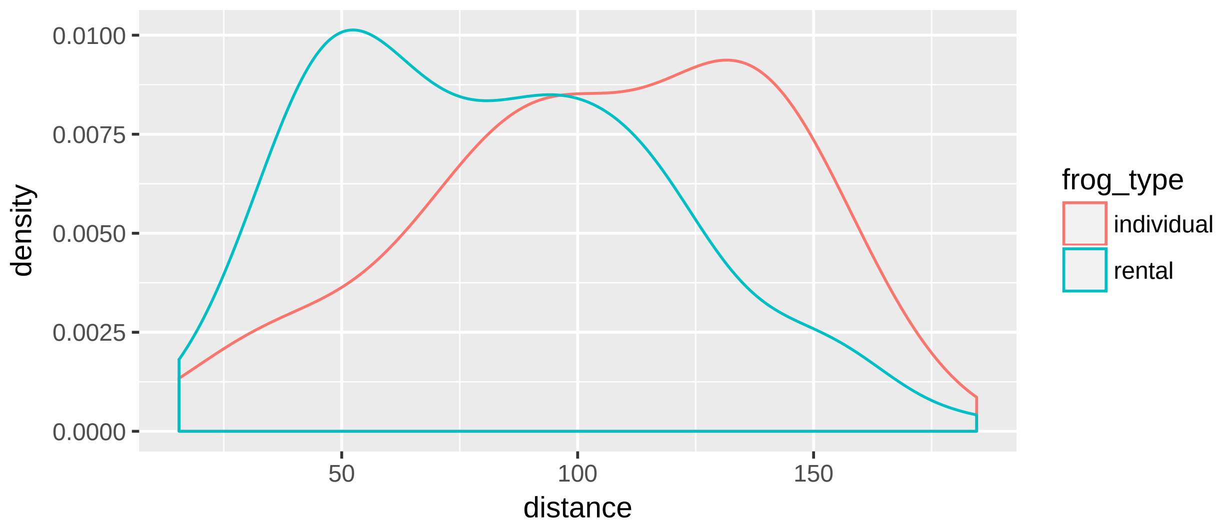
1. Independence within each group

2. Independence across groups

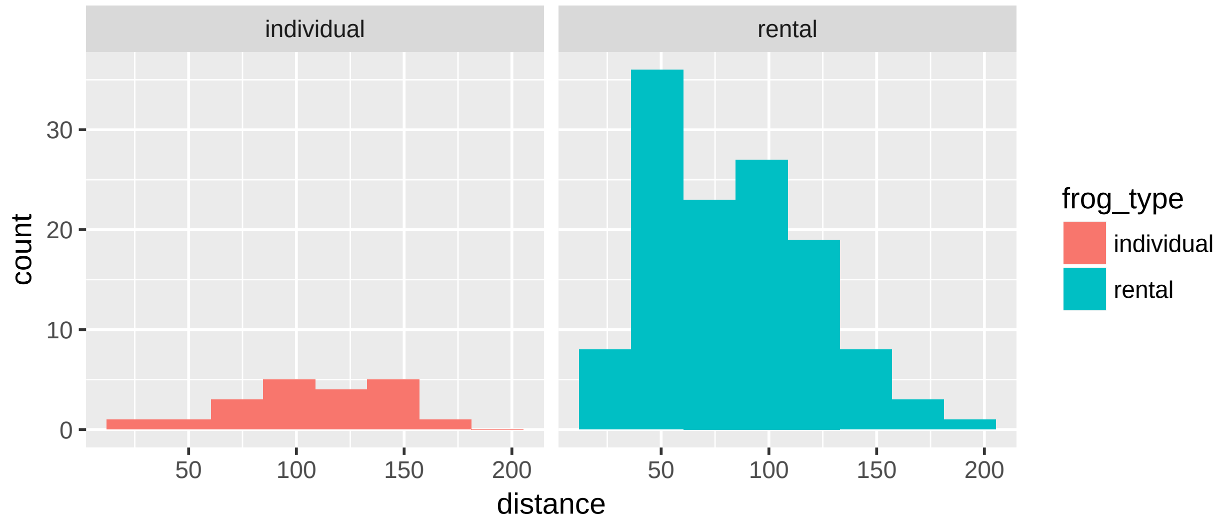
3. Nearly normal distribution

To deal with the issue of independence within each group, let's look at just one of the jumps – how about the second. Here's a plot of jump distance by frog type for just the second jump:

```r
frogs_amateur_jump_2 <- frogs_amateur %>% filter(jump_n == 2)
ggplot() +
  geom_density(mapping = aes(x = distance, color = frog_type),
    data = frogs_amateur_jump_2)
```




```r
ggplot() +
  geom_histogram(mapping = aes(x = distance, fill = frog_type),
    bins = 8,
    data = frogs_amateur_jump_2) +
  facet_wrap(~ frog_type)
```



4. Sample size

```r
table(frogs_amateur_jump_2$frog_type)
```

```
##
## individual        pro     rental    unknown
##         20          0        125          0
```

Calculate a p-value

```r
frogs_rental_jump_2 <- filter(frogs_amateur_jump_2, frog_type == "rental")
frogs_individual_jump_2 <- filter(frogs_amateur_jump_2, frog_type == "individual")

t.test(
  frogs_rental_jump_2$distance,
  frogs_individual_jump_2$distance,
  alternative = "two.sided",
  conf.level = 0.95)
```

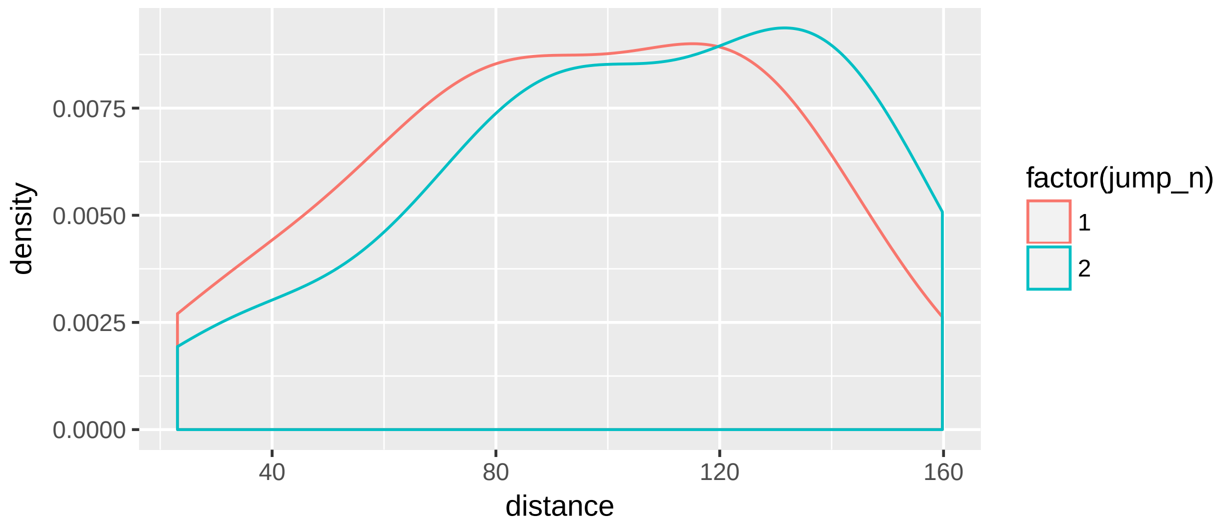
```
##
##  Welch Two Sample t-test
##
## data:  frogs_rental_jump_2$distance and frogs_individual_jump_2$distance
## t = -2.6071, df = 25.33, p-value = 0.01509
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -41.264804  -4.855552
## sample estimates:
## mean of x mean of y
##  82.28787 105.34805
```

Draw a conclusion for the hypothesis test.

Find a 95% confidence interval for the difference in means and interpret it in context.

Example 2: Are mean jump distances the same for the first and second jumps, among frogs brought by individuals?

```
frogs_individual_jumps_1_2 <- frogs_amateur %>%
  filter(frog_type == "individual" & jump_n >= 1 & jump_n <= 2)
ggplot() +
  geom_density(mapping = aes(x = distance, color = factor(jump_n)), data = frogs_individual_jumps_1_2)
```



State Null and Alternative Hypotheses

Check Assumptions for Two-Sample t test

1. Independence within each group

2. Independence across groups

3. Nearly normal distribution

4. Sample size

We can't do a regular two-sample t test because assumptions of independence across groups are violated. Instead, we can do a "paired t test":

- Calculate differences between observed values for each pair
- Perform a t test of whether the average difference is equal to 0

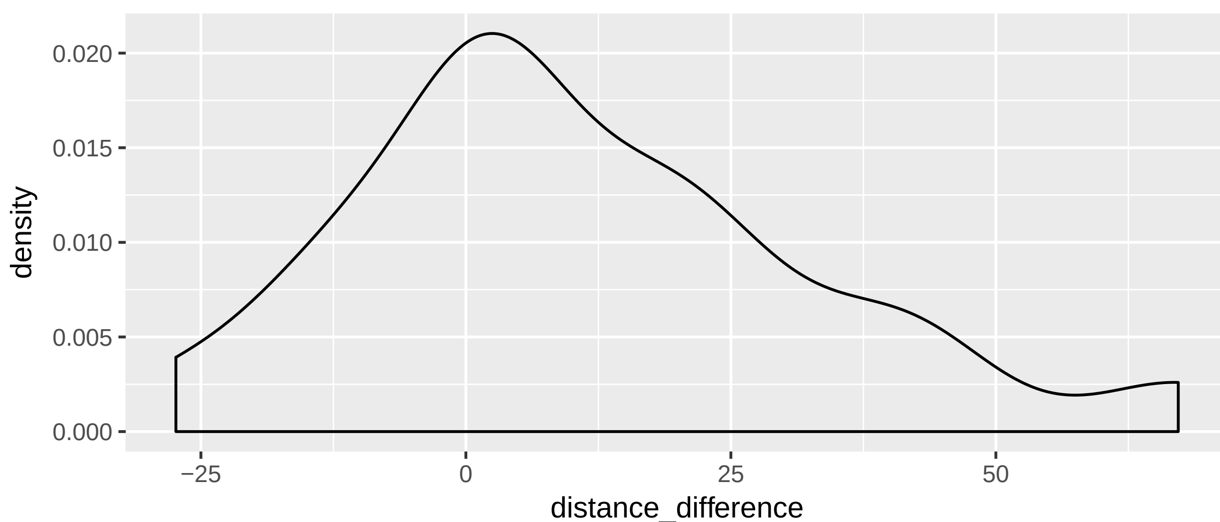
```
frogs_individual_wide <- frogs_individual_jumps_1_2 %>%
  select(id, frog_type, distance, jump_n) %>%
  spread(key = jump_n, value = distance, sep = "_distance_") %>%
  mutate(distance_difference = jump_n_distance_2 - jump_n_distance_1)

head(frogs_individual_wide)
```

```
##    id  frog_type jump_n_distance_1 jump_n_distance_2 distance_difference
## 1 332 individual           136.460           137.070               0.610
## 2 339 individual            83.692            85.206               1.514
## 3 354 individual            57.835            88.758              30.923
## 4 355 individual           118.870           121.460               2.590
## 5 356 individual            64.332           131.540              67.208
## 6 357 individual            86.529                NA                  NA
```

```
ggplot() +
  geom_density(mapping = aes(x = distance_difference), data = frogs_individual_wide)
```

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



Check Assumptions for *Paired* t-test

1. Differences between paired observations are independent across different pairs

2. Differences between paired observations follow a nearly normal distribution

3. Sample size

Calculate a p-value (shown two ways, you only have to do one)

```
t.test(
  frogs_individual_wide$jump_n_distance_2,
  frogs_individual_wide$jump_n_distance_1,
  mu = 0,
  alternative = "two.sided",
  paired = TRUE
)
```

```
##
##  Paired t-test
##
## data:  frogs_individual_wide$jump_n_distance_2 and frogs_individual_wide$jump_n_distance_1
## t = 2.2044, df = 19, p-value = 0.04002
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.5596879 21.5895121
## sample estimates:
## mean of the differences
##                 11.0746
```

```
t.test(
  frogs_individual_wide$distance_difference,
  mu = 0,
  alternative = "two.sided"
)
```

```
##
##  One Sample t-test
##
## data:  frogs_individual_wide$distance_difference
## t = 2.2044, df = 19, p-value = 0.04002
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   0.5596879 21.5895121
## sample estimates:
## mean of x
##   11.0746
```

Draw a conclusion for the hypothesis test.

Find a 95% confidence interval for the difference in means and interpret it in context.