Lecture 10: More Simulations

STAT GR5206 *Statistical Computing & Introduction to Data Science*

Linxi Liu
Columbia University

November 9, 2018

Last Time

- **Pseudo-random Number Generators**. What makes numbers 'random' and the Linear Congruential Estimator.
- **Simulating Random Variables in** `R`: Built-in `R` functions for common distributions and the `sample()` function.
- **Simulating Random Variables from Uncommon Distributions**. Use the Accept-Reject Algorithm.

This Time

- **Permutation Tests**. Are two distributions the same?
- **More Examples**...

Permutation Tests

Cats

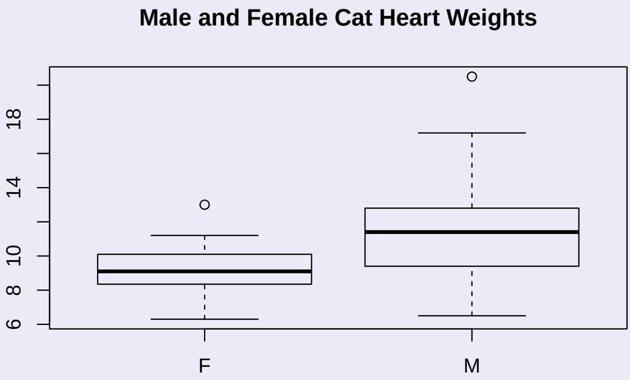
The cats dataset includes the heart and body weights of samples of male and female cats. All the cats are adults and over 2 kg in body weight.

```
> # install.packages("MASS")
> library(MASS)
> head(cats)
```

```
  Sex Bwt Hwt
1   F 2.0 7.0
2   F 2.0 7.4
3   F 2.0 9.5
4   F 2.1 7.2
5   F 2.1 7.3
6   F 2.1 7.6
```

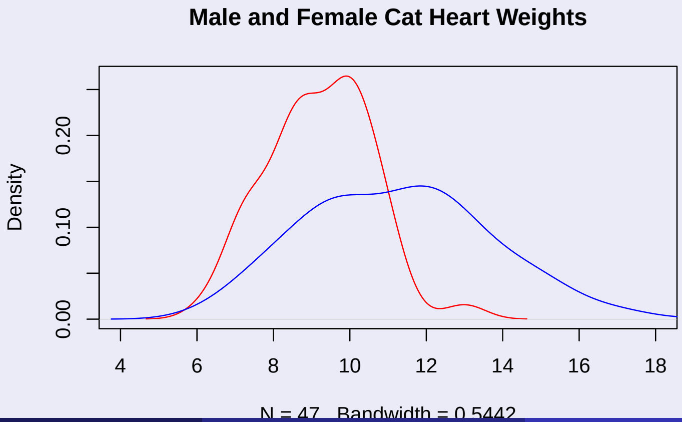
Male and Female Cat Heart Weights

```
> boxplot(cats$Hwt ~ cats$Sex,
+         main = "Male and Female Cat Heart Weights")
```



Male and Female Cat Heart Weights

```
> plot(density(cats$Hwt[cats$Sex == "F"]), col = "red",
+      xlim = c(4, 18), main = "Male and Female Cat Heart Weights")
> lines(density(cats$Hwt[cats$Sex == "M"]), col = "blue")
```



Male and Female Cat Heart Weights

Recall the Independent Two-Sample T-test

- Tests if the population means of two samples are equal.
- Assume the data in each sample come from a normal distribution.

```
> girlcats <- cats$Sex == "F"
> t.test(cats$Hwt[girlcats], cats$Hwt[!girlcats])
```

```
        Welch Two Sample t-test

data:  cats$Hwt[girlcats] and cats$Hwt[!girlcats]
t = -6.5179, df = 140.61, p-value = 1.186e-09
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -2.763753 -1.477352
sample estimates:
mean of x mean of y
 9.202128 11.322680
```

Male and Female Cat Heart Weights

What if the assumptions don't hold?

- In our dataset, 47 female cats and 97 male cats.
- We study test statistic:

$$\hat{D} = mean(X_F) - mean(X_M),$$

  where X_F are female cat heart weights and X_M are male cat heart weights.
- For our data $\hat{D} = -2.12$.

Male and Female Cat Heart Weights

What if the assumptions don't hold?

- In our dataset, 47 female cats and 97 male cats.
- We study test statistic:

$$\hat{D} = mean(X_F) - mean(X_M),$$

where X_F are female cat heart weights and X_M are male cat heart weights.

- For our data $\hat{D} = -2.12$.

Permutation Principle

If there were no difference in male and female heart weights (null hypothesis), then all datasets obtained by randomly assigning 47 of the values in `cats$Hwt` to female cats and 97 to male cats would have equal chance of being observed in the study.

Male and Female Cat Heart Weights

What if the assumptions don't hold?

This gives us

$$\binom{144}{47} = \frac{144!}{47!\,97!} = 2.231033 \times 10^{38}$$

possible two-sample datasets (under the null hypothesis).

How rare is our result?

How many of these datasets have $|mean(X_F) - mean(X_M| \geq 2.12$? What is the probability of seeing differences in the group means as extreme (or more extreme) than ours?

Male and Female Cat Heart Weights

What if the assumptions don't hold?

- Could consider each of the 2.231033×10^{38} possible two-sample datasets, calculate $|mean(X_F) - mean(X_M)|$ for each, and then compare to our original $\hat{D}$...
- Instead we use a **permutation test**, which randomly samples from the 2.231033×10^{38} possible two-sample datasets and estimates a p-value based on our original $\hat{D}$.

Permutation Test

Steps

For step $i = 1, 2, \ldots, P$,

- Randomly assign heart weights in `cats$Hwt` as either male or female with exactly 97 males and 47 females.
- Compute $\hat{D}_i = |mean(X_{F,i}) - mean(X_{M,i})|$ where $X_{F,i}$ are female cat heart weights in step i and $X_{M,i}$ are male cat heart weights.

Finally, we calculate a (two-sided) p-value as follows:

$$\frac{1}{P}\sum_{i=1}^{P}\mathbb{I}(|\hat{D}_i| \geq |\hat{D}|).$$

Permutation Test

What are we actually testing?

- Using the permutation test, we test the null hypothesis that the two samples are from the same distribution.
- How do we do this? Have we seen enough evidence (in the form of the observed difference between the sample means being large enough) to reject the null hypothesis that the two groups have identical probability distributions?

Check Yourself: Permutation Test

Task

Fill in the following code to run a permutation test on the cats$Hwt data.

```
> girlcats <- cats$Sex == "F"
> Dhat     <- mean(cats$Hwt[girlcats])
>                - mean(cats$Hwt[!girlcats])
> nf   <- sum(girlcats); nm <- sum(!girlcats)
> P    <- 10000
> sample_diffs <- rep(NA, P)
> for (i in 1:P) {
+   ###################
+   ## Add code here ##
+   ###################
+ }
> pval <- mean(abs(sample_diffs) >= abs(Dhat))
```

Check Yourself: Permutation Test

Solution

```
> girlcats <- cats$Sex == "F"
> Dhat <- mean(cats$Hwt[girlcats])-mean(cats$Hwt[!girlcats])
> nf   <- sum(girlcats); nm <- sum(!girlcats)
> P    <- 10000
> sample_diffs <- rep(NA, P)
> for (i in 1:P) {
+   perm_data <- cats$Hwt[sample(1:(nf+nm))]
+   meanf     <- mean(perm_data[1:nf])
+   meanm     <- mean(perm_data[-(1:nf)])
+   sample_diffs[i] <- meanf - meanm
+ }
> pval <- mean(abs(sample_diffs) >= abs(Dhat))
> pval
```

```
[1] 0
```