Modeling and Simulation in R

Randall Pruim

Calvin College

Lady Tasting Tea

Lady Tasting Tea

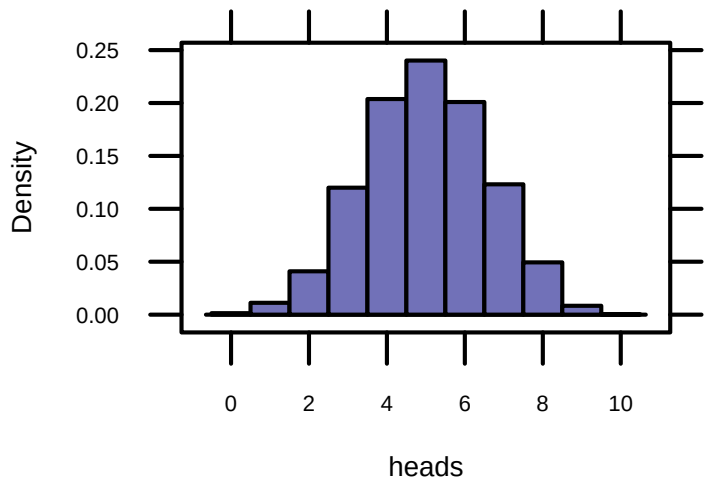
```
> rflip()


Flipping 1 coins [ Prob(Heads) = 0.5 ] ...

H

Result: 1 heads.


> rflip(10)


Flipping 10 coins [ Prob(Heads) = 0.5 ] ...

T T H T T H H T T H

Result: 4 heads.
```

Lady Tasting Tea

```
> do(3) * rflip(10)


   n heads tails
1 10     4     6
2 10     6     4
3 10     5     5


> lady <- do(5000) * rflip(10)
> tally(~heads, data = lady)



   0    1    2    3    4    5    6    7    8    9   10
   7   56  205  600 1019 1201 1005  616  247   42    2
Total
 5000
```

Lady Tasting Tea

```
> xhistogram(~heads, data = lady, width = 1)
```



Are Boys' Feet Bigger Than Girls'?

```
> mean(length ~ sex, KidsFeet)


    B     G
25.11 24.32
```

The mosaic function `mm()` provides a different way of comparing these means.

```
> mm(length ~ sex, KidsFeet)


Groupwise Model.
Call:
length ~ sex

Coefficients:
   B     G
25.1  24.3
```

Are Boys' Feet Bigger Than Girls'?

Let's see how much bigger the boys' feet are (on average):

```
> lengthDiff <- diff(mean(length ~ sex, KidsFeet))
> lengthDiff


      G
-0.7839
```

Q: Is -0.7839 a big difference?

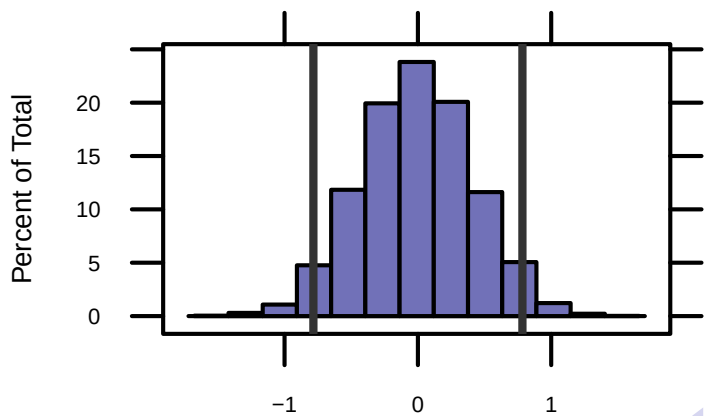
A: Let's compare to random data generated by shuffling the `sex` labels.

```
> do(2) * mm(length ~ shuffle(sex), KidsFeet)


      B     G sigma r.squared
1 24.64 24.81 1.333  0.003794
2 24.82 24.63 1.332  0.005258
```

Are Boys' Feet Bigger Than Girls'?

Let's see how our data compare with data obtained by shuffling the `sex` labels.

```
> Feet <- do(5000) * mm(length ~ shuffle(sex), KidsFeet)
> Feet <- transform(Feet, D = G - B)
> histogram(~D, data = Feet)
> ladd(panel.abline(v = lengthDiff))
> ladd(panel.abline(v = -lengthDiff))
> tally(~abs(D) >= abs(lengthDiff), data = Feet, format = "proportion")



  TRUE  FALSE  Total
0.0616 0.9384 1.0000
```



Are Boys' Feet Bigger than Girls'?

Linear model approach

```
> model <- lm(length ~ sex, data = KidsFeet)
> summary(model)



Call:
lm(formula = length ~ sex, data = KidsFeet)

Residuals:
   Min     1Q Median     3Q    Max
-2.721 -0.713 -0.121  0.795  2.395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.105      0.285   88.18   <2e-16 ***
sexG          -0.784      0.408   -1.92    0.062 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.27 on 37 degrees of freedom
Multiple R-squared: 0.0908, Adjusted R-squared: 0.0662
F-statistic: 3.69 on 1 and 37 DF,  p-value: 0.0623
```

Are Boys' Feet Bigger than Girls'?

The traditional t-test approach:

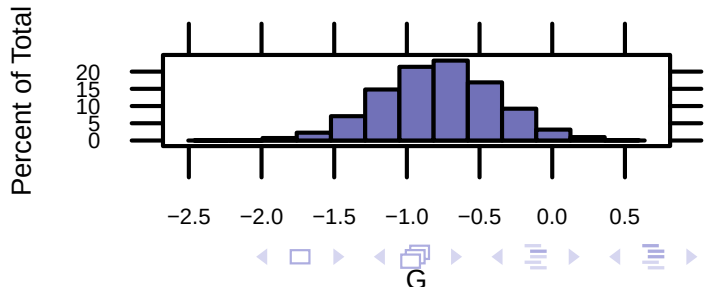
```
> t.test(length ~ sex, data = KidsFeet)



Welch Two Sample t-test

data:  length by sex
t = 1.917, df = 36.27, p-value = 0.06308
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04502  1.61292
sample estimates:
mean in group B mean in group G
          25.11           24.32



> pval(t.test(length ~ sex, data = KidsFeet, equal.var = TRUE))


p.value
0.06308
```

How Much Bigger Are Boys' Feet than Girls'?

Resampling approach:

```
> mean(length ~ sex, data = resample(KidsFeet))


    B     G
24.84 23.99


> diff(mean(length ~ sex, data = resample(KidsFeet)))


     G
-1.319


> rFeet1 <- do(5000) * diff(mean(length ~ sex, data = resample(KidsFeet)))
```

```
> qdata(c(0.025, 0.975), rFeet1$G)


    2.5%    97.5%
-1.55559 -0.01339
```

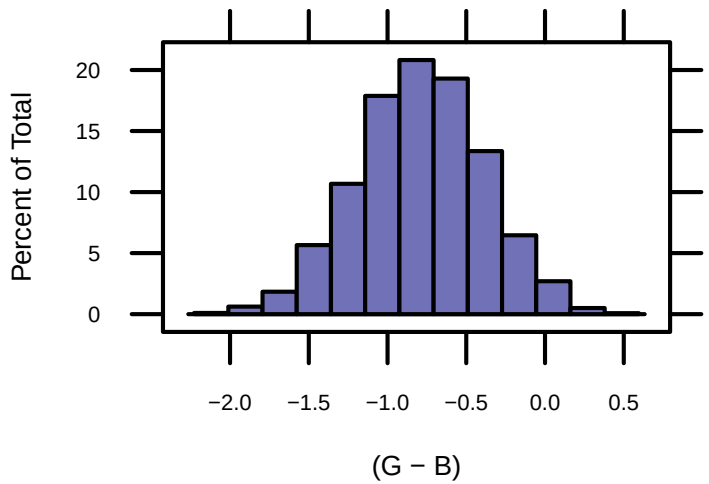


How Much Bigger Are Boys' Feet than Girls'?

Resampling Approach using `mm()`:

```
> rFeet2 <- do(5000) * mm(length ~ sex, data = resample(KidsFeet))
```

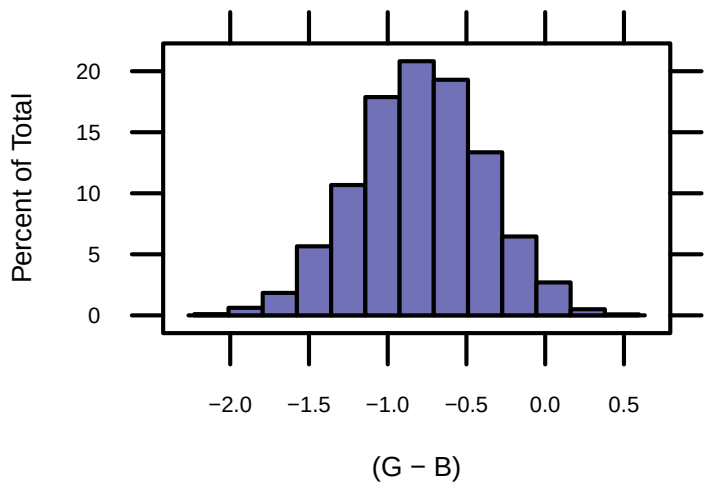
```
> rFeet2 <- transform(rFeet2, D = G - B)
> head(rFeet2, 2)

      B     G sigma r.squared       D
1 24.88 24.34 1.074   0.05944 -0.5345
2 25.03 24.39 1.214   0.06644 -0.6313

> histogram(~(G - B), data = rFeet2)
> qdata(c(0.025, 0.975), rFeet2$D)

    2.5%    97.5%
-1.58400 -0.02329
```



How Much Bigger Are Boys' Feet than Girls'?

Standard Error method:

```
> histogram(~(G - B), data = rFeet2)
> pm <- c(-1, 1)
> lengthDiff + pm * 2 * sd(~D, data = rFeet2)


[1] -1.58322  0.01533
```



How Much Bigger Are Boys' Feet than Girls'?

```
> model <- lm(length ~ sex, data = KidsFeet)
> confint(model)


             2.5 %   97.5 %
(Intercept) 24.53 25.68186
sexG        -1.61  0.04252


> confint(t.test(length ~ sex, data = KidsFeet))


mean in group B mean in group G           lower           upper
       25.10500        24.32105        -0.04502         1.61292
          level
        0.95000
```

The Power of Linear Models

```
> lm( y ~ 1 )       # 1-sample t
> lm( y ~ a )       # 2-sample t, ANOVA
> lm( y ~ a + b )   # 2-way ANOVA (no interaction)
> lm( y ~ a * b )   # 2-way ANOVA (with interaction)
> lm( y ~ x )       # simple linear regression
> lm( y ~ x1 + x2 )# multiple regression (additive)
> lm( y ~ x1 * x2 )# multiple regression (interaction)
> lm( y ~ x + a )   # regression with covariate
```

Other stuff

- More explanatory variables
- Transformations
- Categorical response via logistic regression (using `glm()`)
- Chi-squared test for two-way tables can be replaced by

```
> glm(a ~ b)
```

Two Questions

Q: Should the Intro Stat Course abandon old favorites (t-tests, etc.) and focus on linear models instead?

Q: Should the Intro Stat Course abandon Normal-based inference in favor of simulation and resampling methods?