MORE PRACTICE

Researchers are interested if companies have different sugar content in children cereals than in adults cereals. They randomly collected the sugar content (percentage of weight) on 19 brands of children cereal and 26 brands of adult cereal from a population of companies in order to answer the question of interest. Lets examine histograms of the sugar content by each group to begin checking our assumptions.

```
summary(cereal)
```

```
##      sugar             type
##  Min.   : 1.30   adult   :25
##  1st Qu.: 8.03   children:19
##  Median :20.70
##  Mean   :26.50
##  3rd Qu.:44.58
##  Max.   :60.30
```

```
bwplot(sugar~type,data=cereal)
```

```
histogram(~sugar|type,data=cereal, col = "lightgray")
```

So it appears that the children cereals may be more left-skewed than we would expect from a normal distribution. Later in this class, we will look at some simple ways that we can fix this problem. Stay tuned.

Is our assumption of equal variances safe to assume?

```
favstats(sugar~type,data=cereal)
```

```
##     .group  min    Q1 median    Q3  max  mean    sd  n missing
## 1    adult  1.3  4.40    8.5 16.20 30.2 11.07 7.536 25       0
## 2 children 33.6 43.65   45.9 50.35 60.3 46.80 6.418 19       0
```

What are our hypotheses?

- H_0:
- H_A:
- What is the scope of inference for this example?

- We will now perform the hypothesis test using both the parametric and permutation approaches.

```
t.test(sugar~type,data=cereal,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  sugar by type
## t = -16.58, df = 42, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -40.08 -31.38
## sample estimates:
##    mean in group adult mean in group children
##                  11.07                  46.80
```

- What is the test statistic?

- What is the distribution of the test statistic under the null hypothesis?

- What is the p-value, decision, and conclusion from the parametric test?

- What is the interpretation of the confidence interval *in the context of the problem*? (make sure to check which direction we are working with)

Does our conclusion change if we use the non-parametric approach?

```
set.seed(824862)
bootdist2<-do(1000)*compareMean(sugar~type,data=resample(cereal))
hist(bootdist2$result,col="gray",main="Histogram of bootstrapped Difference in Means", nclass = 50,
             freq = F, xlab = "bootstrapped differences in means (children - adult)")
abline(v=qdata(c(.025,0.975),bootdist2$result)$quantile,col="blue",lwd=3)
```

```
#pull off endpoints of the interval for 95\% CI
qdata(c(.025,0.975),bootdist2$result)

##       quantile     p
## 2.5%     31.74 0.025
## 97.5%    39.97 0.975

#99%CI
qdata(c(.005,0.995),bootdist2$result)

##       quantile     p
## 0.5%     30.30 0.005
## 99.5%    40.98 0.995

observed2<-compareMean(sugar~type,data=cereal)
observed2

## [1] 35.73

nulldist2<-do(5000)*compareMean(sugar~shuffle(type),data=cereal)
hist(nulldist2$result,col="lightgray",main="Histogram of Null Distribution",xlim = c(-38,38), nclass = 50,
        probability = F, labels = T)
abline(v=observed2,lwd=3)
abline(v=-observed2,lwd=3)
```

```
pdata(abs(observed2),abs(nulldist2$result),lower.tail=F)

## [1] 0
```

Does 0 really mean 0?

If you are using a non-parametric test and you find a p-value of 0, that does not mean that there is a 0% chance that you would observe a result as or more extreme than your observed result if the null hypothesis were true. It means that there is less than a $100 * \frac{1}{\#ofsimulations}$*% chance you would observe a result as or more extreme than the one you did if the null hypothesis is true.*