survey Dataset

The Dataset

The *survey* dataset—stored in the *MASS* library—looks at variables from a survey of over 200 students at the University of Adelaide.

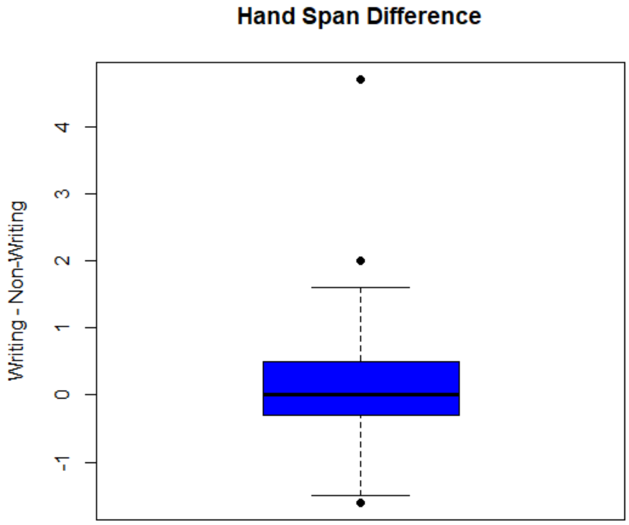
- *sex*: The subject's sex.
- *Wr.Hnd*: The subject's writing hand span (In centimeters).
- *NW.Hnd*: The subject's non-writing hand span (In centimeters).
- *W.Hnd*: The subject's writing hand.
- *Fold*: A factor indicating which arm is on top when the subject folds their arms.
- *Pulse*: The subject's pulse in beats per minute.
- *Clap*: A factor indicating which hand is on top when the subject claps their hands.
- *Exer*: A factor indicating how often the subject exercises.
- *Smoke*: A factor indicating how often the subject exercises
- *Height*: The subject's height in centimeters
- *M.I*: Whether the subject's height was expressed in metric or imperial units/
- *Age*: The student's age in years

Potential Questions

This dataset offers several interesting questions, but also a few challenges as well. Many observations have missing data stored as NA values, meaning that students will have to use the *na.rm* option of functions often. It could additionally be interesting to have students try to impute these values. Another option to work with this dataset is to poll students in a class and compare the Australian students in the dataset with students in the class Regardless, there is much that students can work with just using the data on its own.

1. **Are writing hands and non-writing hands have different spans? (Paired mean)**
2. Do people who smoke at all have higher pulses? (2-sample mean)
3. Do people who exercise at all have lower pulses? (2-sample mean)
4. **Do people who smoke exercise less regularly? (2-sample proportion)**
5. Do people clap with their dominant hands on top? (2-sample proportion)
6. Do people fold their arms with their dominant hand arm on top? (2-sample proportion)
7. Are age and pulse associated? (Correlation)
8. Are height and pulse associated? (Correlation)
9. **Can we predict height with hand span? (Regression)**
10. How do we predict an individual's pulse?*

Question 1



It would seem reasonable to expect that your hands would be the same size. At the same time, your dominant hand tends to get more work than your non-dominant, meaning that your dominant hand might potentially have a little more flexibility to it. Regardless which way you lean, this is something that we can test in this data. It should be noted that since there's overlap in the two samples—each writing/non-writing hand pair come from the same person—this is a paired test, not a two-sample test for means. Looking at the boxplot of the differences, it looks like the difference is going to be zero or close to it, but as we know in hypothesis testing small deviations from the proposed null value may be statistically significant.

In this case, we'll test $H_A : \mu_D > 0$ at the $\alpha = 0.05$ level, where μ_D is the population average difference. The test returns a test statistic of $t = 2.2168$ with a p-value of $P\big(t_{235} > 2.1268\big) = 0.01724$. Thus, we'd reject the null hypothesis and conclude that the population difference between hand span—writing minus non-writing—is greater than zero.

We should note that our sample size assumptions hold here, as well as the rough normality to boot. The boxplot shows one particularly strange outlier, which could in theory have affected their results given how close to 0 the sample average difference is. An interesting question would be does this difference differ for left-handers versus right-handers. A quick calculation of the averages shows some interesting data, though we'd have some challenges regarding to our sample sizes.

```
library(MASS)
data(survey)
boxplot(survey$Wr.Hnd-survey$NW.Hnd,
pch=16,col=''blue'', main=''Hand Span Difference'',
ylab=''Writing - Non-Writing'')
t.test(Wr.Hnd,NW.Hnd,data=survey,
paired=T,alternative="greater")
mean(survey$Wr.Hnd[survey$W.Hnd==''Left''],
na.rm=T)
mean(survey$Wr.Hnd[survey$W.Hnd==''Right''],
na.rm=T)
```

Question 4

It likely wouldn't be surprising if someone told you that one's exercise habits and smoking habits are related. Generally, most would expect that smokers are going to exercise less. We can look at this using our student data, creating the contingency table of *Exer* and *Smoke.*

	No Exercise	Some Exercise	Exercise Frequently
Never Smoke	18	84	87
Smoke Occasionally	3	4	12
Smoke Regularly	1	7	9
Smoke Heavily	1	3	7

These variables have more than two levels, meaning that we'll have to work on adjusting the variables a bit. For the *smoke* variable, we'll create two groups: the non-smoking group and the group that occasionally, regularly, or heavily smokes. The exercise variable will be grouped into the frequent exercisers and the non and sometimes exercisers. This creates the following combined contingency table.

	Exercises Frequently	Exercises Non-frequently
Non-Smoker	87	102
Smoker	28	19

We'll test the hypothesis $H_A : p_{Non-smoker} - p_{Smoker} > 0$, where p is the proportion of frequent exercisers, at the $\alpha = 0.05$ level. The test statistic was $t = -\sqrt{2.7631} = -1.662$ with a p-value of $P\big(N(0,1) > -1.662\big) = 0.9518$. Thus, we'd fail to reject the null hypothesis and conclude it's plausible that smokers and non-smokers exercise frequently at similar rates. The assumptions related to sample size check out, although the smoker sample sizes are a little on the small size.

This problem is an interesting question on how we have to combine groups to make the two-sample test for proportions work. If we instead combine the exercise frequently and sometimes exercise groups, we get a vastly different contingency table, although one that would make implementing a hypothesis test challenging.

	Exercises	Doesn't Exercise
Non-Smoker	171	18
Smoker	42	5

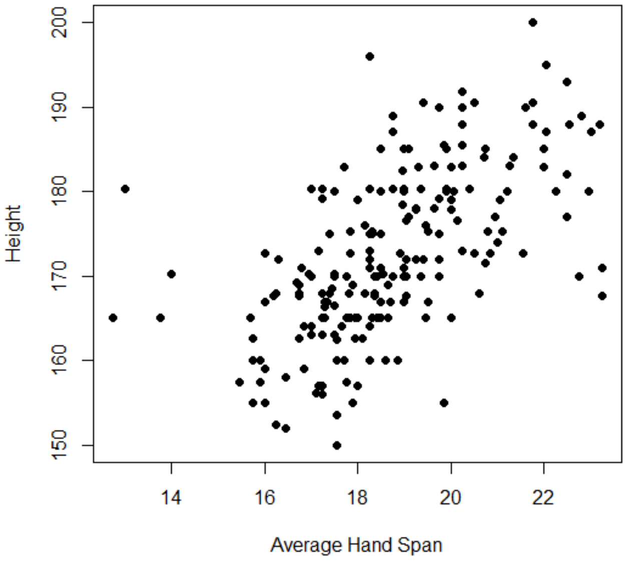
This scenario shows the limitations of the two-sample test for proportions. In this case, to establish an association between smoking status and exercise, we'd need to use the χ^2 test for independence—although that has some assumption problems with this dataset. But seeing how students attack this problem with the limitation of pairwise comparison opens up some interesting possibilities for creating problem solving.

```
library(MASS)
data(survey)
table(survey$Smoke, suvey$Exer)
prop.test(x=c(87,28),n=c(87+102,
28+19),correct=F,
alternative="greater")
-sqrt(2.7631)
```

Question 9

Body proportions are something that we are intuitively familiar with, and something humans have been cognizant of for thousands of years. People are probably most familiar with this concept via Da Vinci's Vitruvian Man. In general though, we expect that, say, hand span will be correlated with height. So a logical extension could be, can we predict height using hand span?

The starting point of this question is what hand span do we want to use? We could choose the larger, the smaller, the average, both—that's a bad idea. In this case, I'll use the average hand span, reasonable in my view as most hands will be similar. Looking at a scatterplot of average hand span and height, we might have a tougher time predicting height than expected.



There's a little more error around the regression line than I might have initially expected, but we can likely still get a decent estimate from a regression. Using the *lm* function, we can find the regression is not bad but not great, with an $R^2 = 0.3576$.

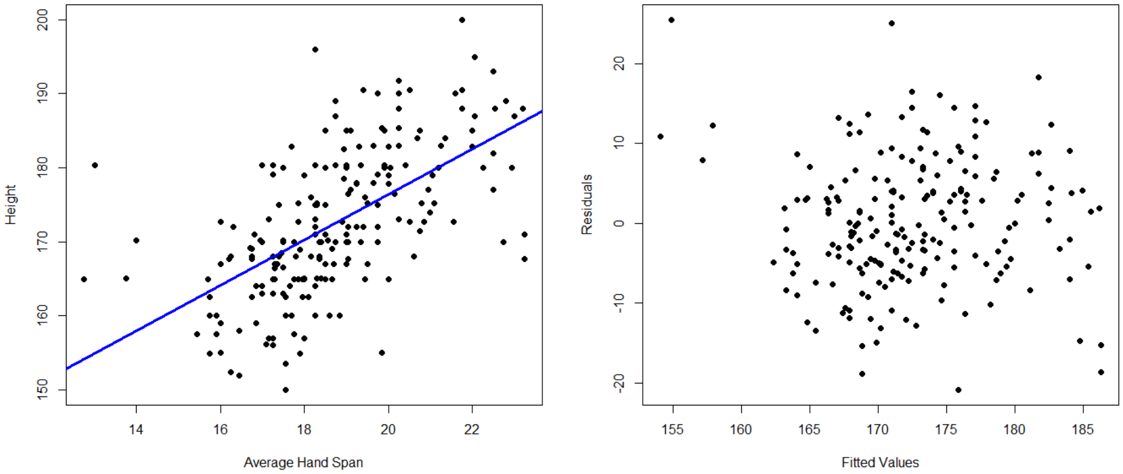
```
Residuals:
     Min       1Q   Median       3Q      Max 
-20.8794  -5.1364  -0.5666   4.3981  25.4735 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 114.9880     5.3884   21.34   <2e-16 ***
Hand          3.0676     0.2865   10.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.931 on 206 degrees of freedom
```

```
  (29 observations deleted due to missingness)
Multiple R-squared:  0.3576,    Adjusted R-squared:  0.3544
F-statistic: 114.6 on 1 and 206 DF,  p-value: < 2.2e-16
```

So, for every centimeter wider the hand span, you can add another 3.0676 centimeters to the expected height. We should note that our residuals confirm that our assumptions appear to be met. Interesting exercises beyond this data could be to see if other measures—wingspan, torso height, inseam, etc.—provide a better prediction of height rather than hand span.



```
library(MASS)
data(survey)
survey$Hand=(survey$Wr.Hnd+survey$NW.Hnd)/2
plot(survey$Hand, survey$Height, pch=16,
xlab=''Average Hand Span'', ylab=''Height'')
lm=lm(Height~Hand, data=survey)
summary(lm)
par(mfrow=c(1,2))
plot(survey$Hand, survey$Height, pch=16,
xlab=''Average Hand Span'', ylab=''Height'')
abline(lm, col=''blue'', lwd=3)
plot(lm$fitted, lm$resid, xlab=''Fitted Values'',
ylab=''Residuals'')
```