Lab 5

September 28, 2016

All these exercises will require the latest PubH6002 library of functions. If you haven't installed it yet, do so. You can install it by going to the packages tab and using Install > Install from Package Archive File, then selecting the file PubH6002_0.24.tar.gz

```
library(PubH6002) # The version should be 0.24
```

If you want the notebook to compile despite errors, remove the # from the beginning of the next line:

```
#allowErrorsInNotebook()
```

Part 1: Numerical variables

This exercise uses the Pima.te dataset from the MASS package. The dataset includes measurement on a population of women of Pima Indian heritage living near Phoenix, Arizona. The women were tested for diabetes according to World Health Organization criteria. The variables in this dataset are:

Variable	Meaning
npreg	number of pregnancies.
glu	plasma glucose concentration
bp	diastolic blood pressure (mm Hg)
skin	triceps skin fold thickness (mm)
bmi	body mass index
ped	diabetes pedigree function
age	age in years
ageGroup	age group in decades
type	Yes or No: diabetic by WHO criteria

The data are in a file called pima.csv. We start by reading it in:

```
pima = read.csv('pima.csv')
```

Next, we "use" the pima dataset so we can easily use the variables inside it

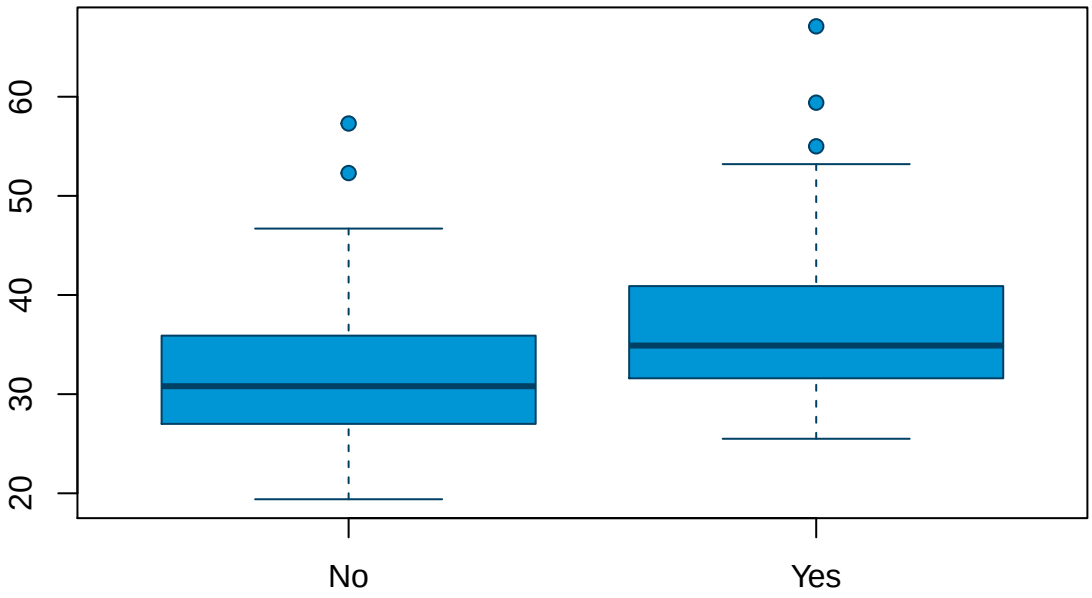
```
use(pima)
```

```
## Note: previous dataset closed. Its variables are not visible anymore.
## Note: the variables in pima are now visible.
```

Question 1

One of the important variables in this dataset is bmi, the body mass index (an indicator of high body fat). Here we are interested in comparing the BMIs of diabebtics (type=Yes) to those of non-diabetics (type=No). Use the boxplot() function to create boxplots of variable bmi by variable type

```
# Given: boxplot(bmi ~ )
boxplot(bmi ~ type)
```



Question 2

Do diabetics (type=Yes) have a higher or lower BMI in the sample?

Answer: higher

Question 3

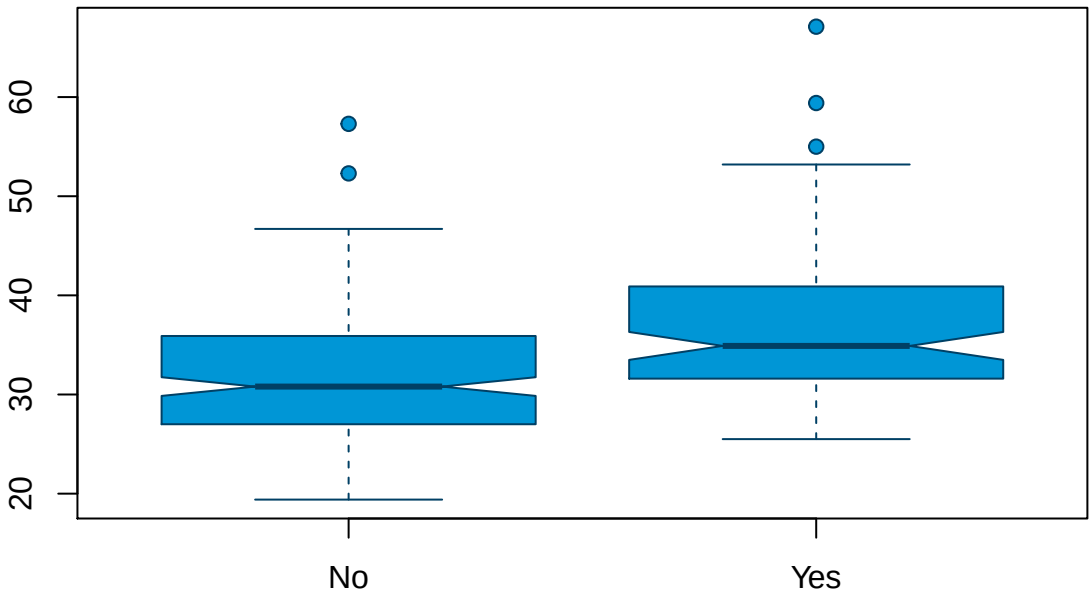
Looking at the boxplots, do they seem about symmetric around the median or skewed?

Answer: somewhat skewed in the Yes group. The No group is about symmetric, or maybe very slightly skewed

Question 4

We are interested in whether BMI values differ between diabetics and non-diabetics in the population, not just in the sample. Add notches to the boxplots to visually assess this

```
# Given: boxplot( , notch=TRUE)
boxplot(bmi ~ type, notch=TRUE)
```



Question 5

Do the notches overlap? What do you conclude about the difference in median BMIs in the population?

Answer: the notches don't overlap. We conclude that the medians appear to be different in the population as well.

Question 6

If the notches did overlap, what could you conclude about the medians of BMI in the population?

Answer: If the notches overlap, there is no clear conclusion. We need to do a formal test to conclude.

Question 7

We are interested in formally testing whether the diabetics and non-diabetics have different mean BMIs in the population. Perform a t test for a difference in means.

```
# Given: t.test( ~ )
t.test(bmi ~ type)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by type
## t = -5.7896, df = 193.96, p-value <0.0001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.532946 -3.212921
## sample estimates:
##  mean in group No mean in group Yes
##          31.63991          36.51284
```

Question 8

From the output, what are the mean BMIs in the diabetic and non-diabetic groups in the sample? Is this consistent with the conclusions about medians from the boxplots?

Answer: 31.64 among non-diabetics, 36.51 among diabetics, so diabetics have a higher mean, which is consistent with their having a higher median BMI in the boxplots

Question 9

From the output, what is the p-value? is it less than 0.05? Do you conclude that such a difference in means is (usual|unusual) if the populations means of BMI are the same?

Answer: yes, it's actually <0.0001. So it would be very unusual to see such a difference in the sample if the population means are the same.

Question 10

Perform a permutation test instead of a t-test.

```
# Given: permutation.test( ~ )
permutation.test(bmi ~ type)
```

```
##
##  Permutation test for a difference in means
##
## data:  bmi by type
## Sample difference = -4.8729, permutations = 1000, p-value = 0.0010
## alternative hypothesis: true difference in means is not equal to 0
## sample estimates:
##  mean in group No mean in group Yes
##          31.63991           36.51284
```

Question 11

From the output, what are the mean BMIs for diabetics and non-diabetics in the sample (same as above, hopefully)? What's the difference between them?

Answer: 31.64 among non-diabetics, 36.51 among diabetics. The difference between them is -4.8729.

Question 12

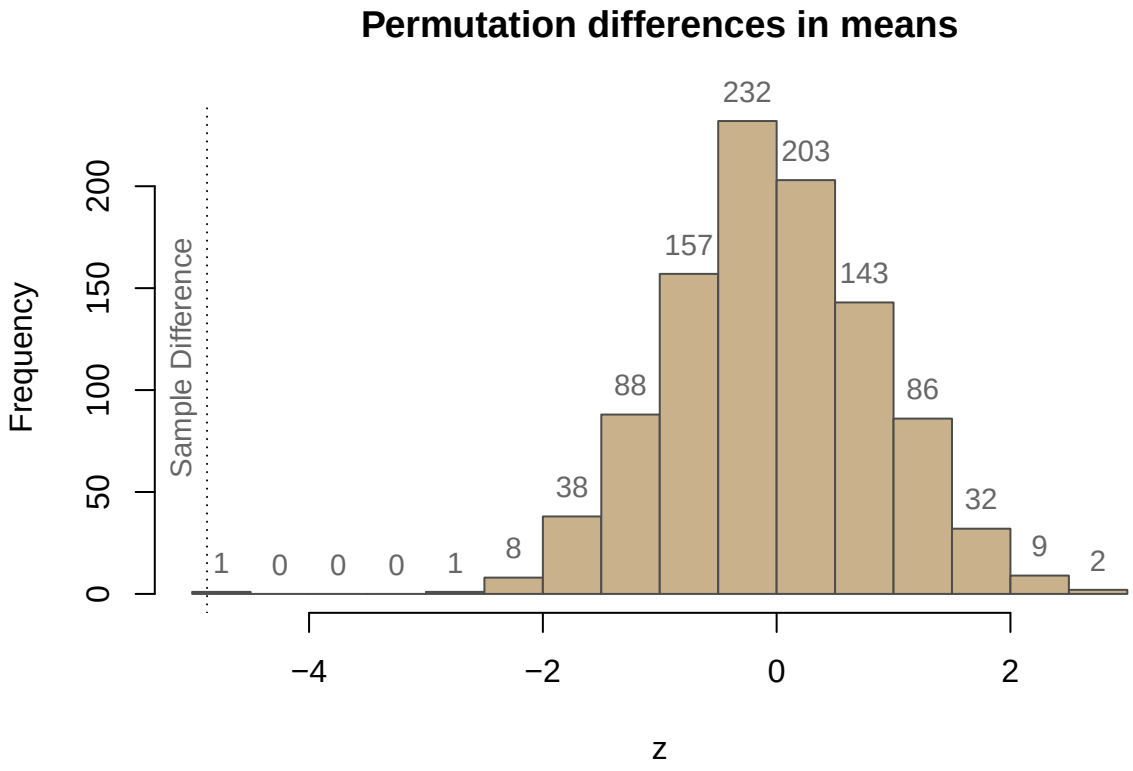
If there is no difference in mean BMI between diabetics and non-diabetics in the population, how likely are we to see the sort of difference we're seeing in the sample? Do you conclude that this sort of difference in the sample is (usual|unusual) if the population BMIs are the same?

Answer: the probability is 0.001, i.e. 1/1000, which is pretty small. A difference of -4.8729 is pretty unusual in the sample if the BMIs are the same in the population

Question 13

Repeat the permutation test above, but use plot=TRUE to display the permutation differences and the observed difference in the sample. Does the sample value look unusual compared to the typical values one expects to see if there are no differences in mean BMI in the population?

```
# Given: permutation.test( ~ , plot=TRUE)
permutation.test(bmi ~ type, plot=TRUE)
```




```
##
##  Permutation test for a difference in means
##
## data:  bmi by type
## Sample difference = -4.8729, permutations = 1000, p-value = 0.0010
## alternative hypothesis: true difference in means is not equal to 0
## sample estimates:
##  mean in group No mean in group Yes
##          31.63991           36.51284
```

Answer: yes, it's much smaller than the differences expected when there are no differences

Question 14

The previous tests looked at the means. Since the distribution of BMI is maybe a little skewed in each group, it would be interesting to compare medians. Use a permutation test to compare the population medians.

```
# Given: permutation.test( ~ , statistic=median)
permutation.test(bmi ~ type, statistic=median)
```

```
##
##  Permutation test for a difference in medians
##
## data:  bmi by type
```

```
## Sample difference = -4.1, permutations = 1000, p-value = 0.0010
## alternative hypothesis: true difference in medians is not equal to 0
## sample estimates:
##  median in group No median in group Yes
##                30.8                34.9
```

Question 15

What are the sample medians in diabetics and non-diabetics? What's the difference between the two medians? If the population medians are the same, what is the probability of seeing a difference this extreme in the sample? If the population medians are the same, would this be a usual|unusual sample difference?

Answer: The medians are 30.8 and 34.9, so they differ by -4.1. If the medians are the same in the distribution, the probability of seeing a difference as extreme as -4.1 is about 0.001, which is small. That makes -4.1 an unusual sample difference if the population medians are the same.

Question 16

One question of interest to the investigators is whether diabetes affects the variability of BMI's, not just their values. One way to test this is to look at whether the standard deviation of BMI is about the same in among diabetics and non-diabetics. Use a permutation test to compare standard deviations between diabetics and non-diabetics

```
# Given: permutation.test(, statistic=sd)
permutation.test(bmi ~ type, statistic=sd)
```

```
##
##  Permutation test for a difference in sds
##
## data:  bmi by type
## Sample difference = -0.80953, permutations = 1000, p-value =
## 0.3220
## alternative hypothesis: true difference in sds is not equal to 0
## sample estimates:
##  sd in group No sd in group Yes
##        6.648015        7.457548
```

Question 17

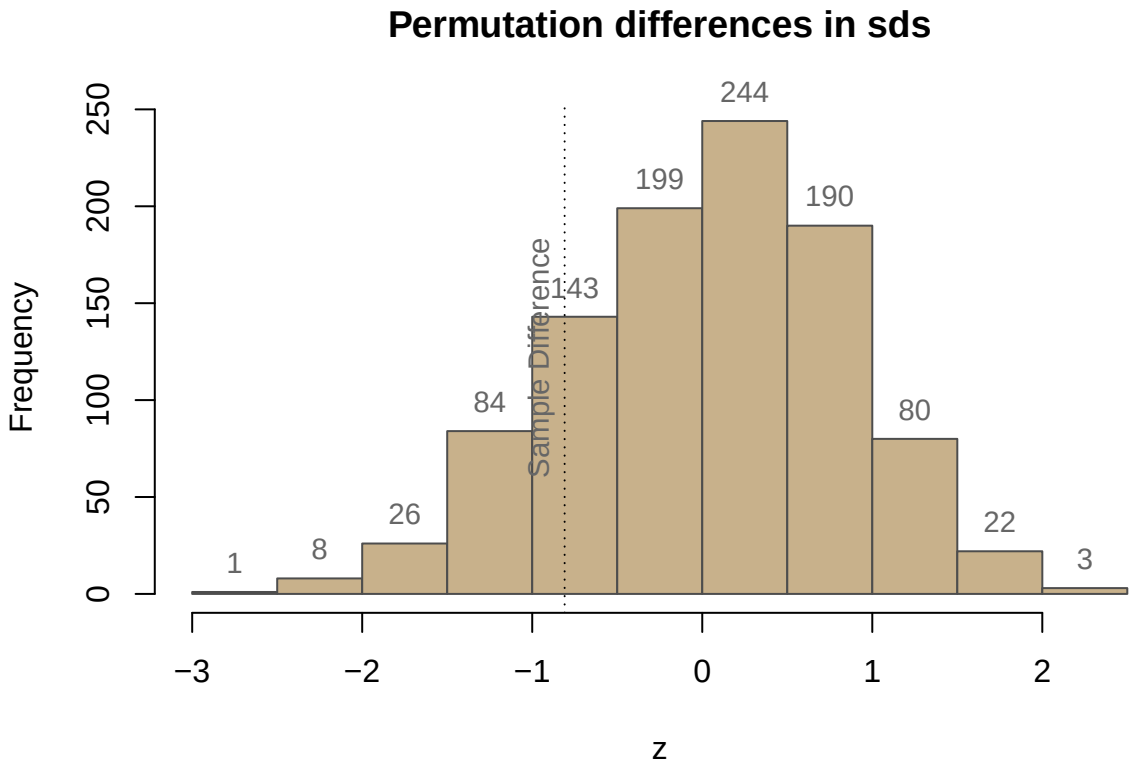
What are the sample standard deviations in diabetics and non-diabetics? What's the difference between these two sd? If the population standard deviations are the same, what is the probability of seeing a difference this extreme in the sample? Is the sample difference unusual?

Answer: the standard deviations are 6.65 among non-diabetics and 7.46 among diabetics, so a difference of -0.80953. If the standard deviation of BMI is the same in both groups in the population, the probability of seeing a difference like -0.80953 is 0.335, i.e. 33.5%, which is not small at all. So if the population standard deviation are the same, -0.80953 is a pretty typical value

Question 18

Add plot=TRUE to the permutation test to visually assess how "typical" the sample difference in standard deviations is. Compared to the sort of differences you'd expect to see if diabetics and non-diabetics have the same standard deviation for BMI in the population, does the sample difference look unusual? Do the conclusions of the previous question seem reasonable?

```
# Given: permutation.test(, statistic=sd, plot=TRUE)
permutation.test(bmi ~ type, statistic=sd, plot=TRUE)
```



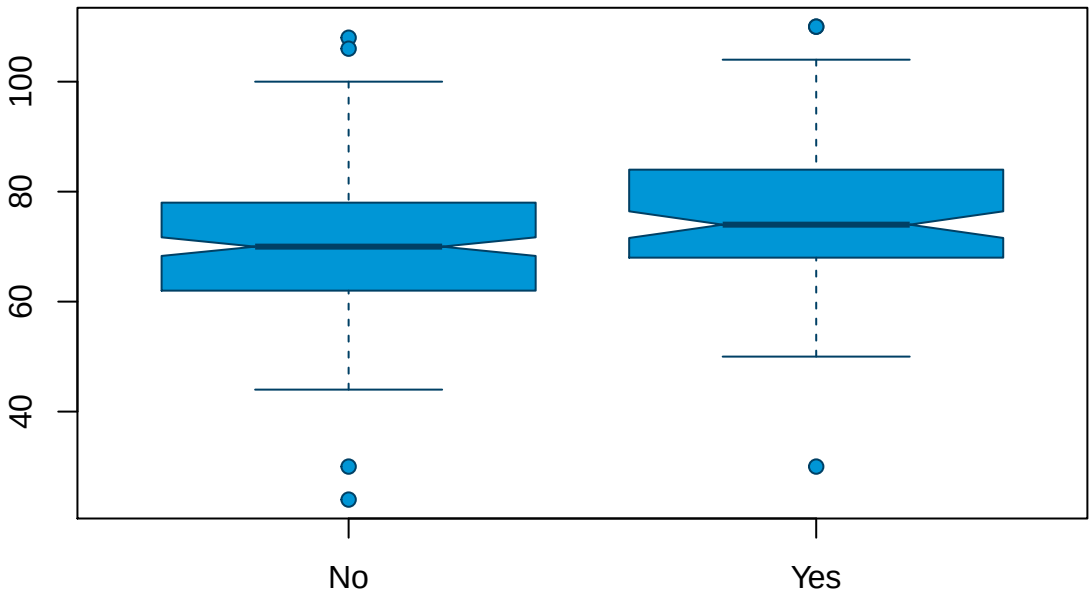

```
##
##  Permutation test for a difference in sds
##
## data:  bmi by type
## Sample difference = -0.80953, permutations = 1000, p-value =
## 0.3220
## alternative hypothesis: true difference in sds is not equal to 0
## sample estimates:
##  sd in group No sd in group Yes
##         6.648015         7.457548
```

Answer: yes, the sample difference looks pretty typical

Question 19

We now look at blood pressure. Produce notched boxplots for blood pressure (variable bp) as a function of diabetes (variable type)

```
# Given: boxplot(, notch=TRUE)
boxplot(bp ~ type, notch=TRUE)
```



Question 20

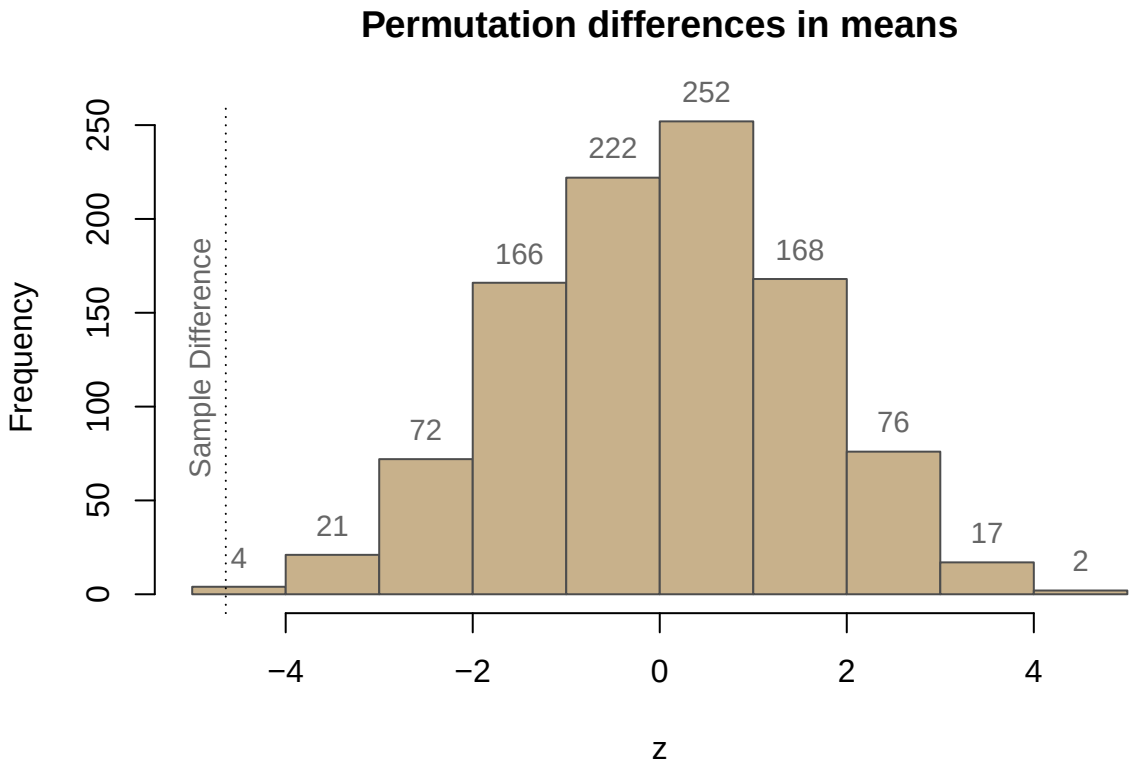
Do the notches seem to overlap, or are they clearly non-overlapping? What do you conclude about the difference in blood pressures in the population?

Answer: they seem to overlap. So no conclusion about blood pressures in the population

Question 21

Do a permutation test comparing mean blood pressures

```
# Given: permutation.test( ~ , plot=TRUE)
permutation.test(bp ~ type, plot=TRUE)
```




```
##
##  Permutation test for a difference in means
##
## data:  bp by type
## Sample difference = -4.6406, permutations = 1000, p-value = 0.0020
## alternative hypothesis: true difference in means is not equal to 0
## sample estimates:
##  mean in group No mean in group Yes
##          70.13004           74.77064
```

Question 22

What are the mean blood pressures in the sample? What the difference between these means?

Answer: 70.13 and 74.77. These differ by -4.6406.

Question 23

If diabetics and non-diabetics have the same mean blood pressure in the population, what is the probability of seeing a difference of -4.6406 or more in the sample? Is the sample difference unusual?

Answer: If there is no difference in the population, -4.6406 should happen with probability 0.003 or 3/1000, so this would be an unusual sample difference in means if there is no difference in means in the population

The End