Final Practice Problems

Mixed practice.

1. How do you calculate the median of a dataset? (Please provide case-by-case scenarios.)

1. Even obs. Take mean of two middles (ordered)
2. odd obs. Take the middle of the ordered data

2. Below are several datasets. Please match the labels to the datasets depending on what type of test you could potentially use on them.

Labels

Permutation Test
χ^2 test for independence
χ^2 Goodness-of-Fit test
Two sample t-test
One sample z-test

Dataset 1.

```
##   kings warriors lakers
## 1     5       20     17
```

χ^2 Goodness of Fit

IRL, probably switched

Dataset 2.

```
##                                  stanford berkeley
## african american                      843     421
## asian                                9990    1476
## chicano/latino                       3192    1125
## native american/alaska native         186      70
## pacific islander                       65      70
## white                                7243    2531
## other/decline to state               1137      70
## mixed                                2525     703
```

χ^2 independence

Dataset 3.

```
##    farenheit  outfit
## 1       53.8 sweater
## 2       57.1 sweater
## 3         56 sweater
## 4       54.6 sweater
## 5       58.7 sweater
## 6       53.8 sweater
## 7       72.9   shirt
## 8       72.2   shirt
## 9       69.5   shirt
## 10      70.6   shirt
## 11        78   shirt
## 12      75.6   shirt
```

Permutation Test

T-test with satisfied assumptions

Dataset 4.

This dataset is similar to Dataset 3 but actually includes 353 more rows that just weren't printed!

two-sample T-test

```
##    farenheit  outfit
## 1       57.7 sweater
## 2       60.7 sweater
## 3       55.9 sweater
## 4         61 sweater
## 5       53.5 sweater
## 6       59.6 sweater
## 7       71.4   shirt
## 8       69.7   shirt
## 9       77.7   shirt
## 10      68.4   shirt
## 11      69.2   shirt
## 12        77   shirt
```

Dataset 5.

For some reason, you know, for sure, the true standard deviation of Chance the Rapper's plays on Spotify. (What are the chances?) You also have these data.

one sample Z-test

```
##    spotify_plays
## 1      307507262
## 2       35403776
## 3      172425446
## 4      166859476
## 5      127965740
## 6       46487943
## 7       17032831
## 8       15322923
## 9      106902309
## 10     114293608
```

3. If you know population standard deviation, then use a ___z___ test! If you don't know the population standard deviation, then use a ___t___ test!
4. Surely, we always want our sample to be representative of our population. We do this by taking a _____ sample. Also, just to be clear, just because the samples are said to be just that, they are not taken haphazardly.

} random, SRS

5. When we want to make a histogram using the library ___ggplot2___, we use the function ___geom_histogram___
6. In dplyr, we use the symbol %>% called the ___pipe operator___ to send our data into functions.

Short Answer.

1. Because assumptions, such as large sample size, may be difficult to satisfy, we have alternative methods. We can evade the large sample size requirement when testing hypotheses by using a _____ test instead of classical testing procedures.

permutation

2. When we construct a 95% confidence interval using the _____ method, we have to find the 2.5th percentile and the 97.5 percentile of our approximate "sampling" distribution to capture ___% of the data within these bounds.

Bootstrap, 95

Santa's Little Statistician.

Directions: It is your job to talk to Santa's elf. Your responses are shown in bold and you need to fill in the blanks.

Welcome to Santa's Workshop! We really needed some help so we decided to invite you, a cheery statistician, to the North Pole! This year, we're working on sending good folk as many presents as possible! That being said, Santa himself said we could only use 15 total presents to make a confidence interval about the weights (in pounds) of the presents. Why weight, you say? Well, Santa's sleigh can only carry up to 1000 lbs! What? Did you think the sleigh was magic?

Part a.

Median Weights.

First item on Santa's stats wishlist today: he wants to know if we can test the hypotheses:

H_0: Median weight of presents is 10 lbs. (He carries 100 presents per ride.)
H_1: Median weight of presents is greater than 10 lbs.

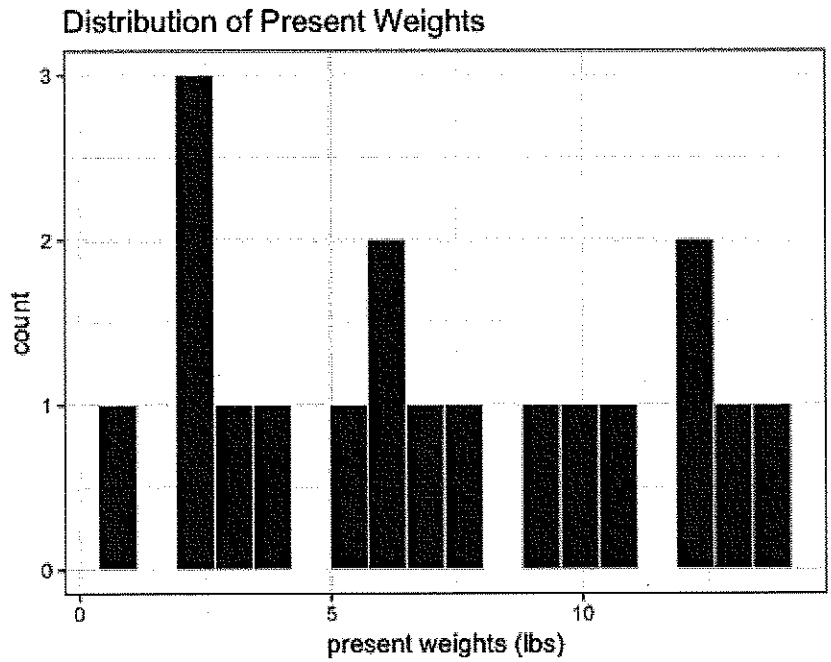
To test the hypothesis, we will use the data below, based on presents chosen random by Santa.

The data frame for the sample of presents:

```
##    present_weights  destinations shipping
## 1                5 north america priority
## 2                9 south america  economy
## 3                6        africa priority
## 4                8          asia  economy
## 5                3        europe priority
## 6                2     australia  economy
## 7                2        europe priority
## 8                7 north america  economy
## 9               10 south america priority
## 10              12        africa  economy
## 11              14          asia priority
## 12              13        europe  economy
## 13               4     australia priority
```

```
## 14          2        europe  economy
## 15         12 north america priority
## 16         11 south america  economy
## 17          6        africa priority
## 18          1          asia  economy
```

A histogram of the sample:



So, can we? (Select one by removing the boldface... I urge you to choose the first option so you can actually finish this worksheet).

(a) Yes, we can using Bootstrap.
(b) No, we can't.

x% CI

[Hint: This question relies on your knowledge that a 95% confidence interval can be used to know whether the p-value for the corresponding two-sided test is < 0.05!]

Hypothesis testing with a confidence interval.

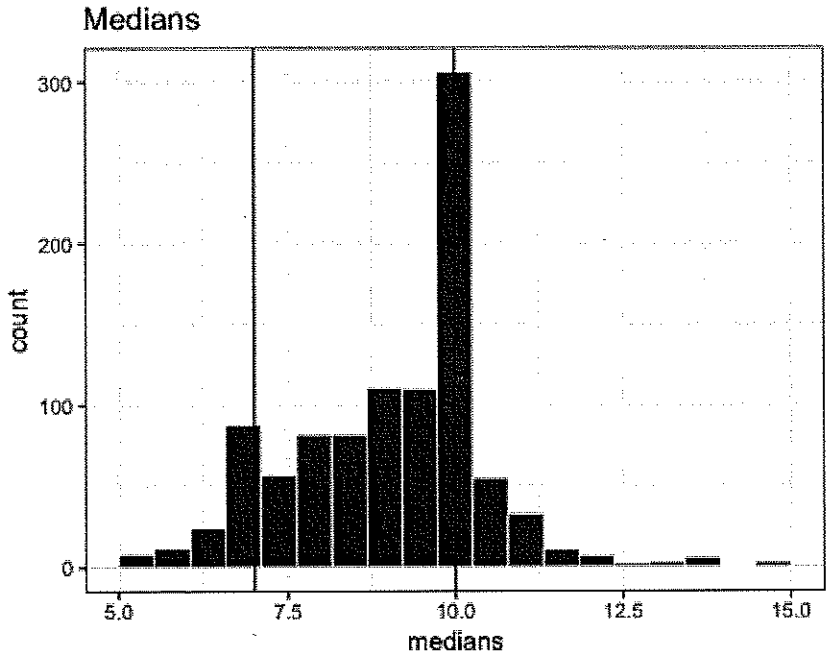
Great! Good to know since we really need to know about these gift weights... Rudolph's back is beginning to hurt! From what I understand about the method you mentioned earlier, we are going to sample from our sample to calculate statistics. Why would we do that?

We sample with **replacement from our sample because we want to estimate the** sampling **distribution of the statistic we are interested in, which is what confidence intervals are normally based off of.**

What was the median of our original present weight dataset?

Our sample's median is 6.5 (numeric value).

Gotcha! Here are the 1000 medians (one for each of the resamples) you asked for in histogram format. I also illustrated the 5th percentile and the 95th percentile using vertical lines.



Here are the values for the 5th and 95th percentiles, which I think you might need to make your confidence interval.

```
median_quantiles
```

```
##   fifth_perc ninetyfifth_perc
## 1          7              11
```

So, what do you think about our original hypotheses? Do you think we need to stop loading our sleigh with 100 presents? (Use as much space as you need!)

Based on our data, we can make a 90% CI. 10 is within our CI.

We fail to reject H_0.

↑ interpret this in context!

Part b.

Weights between shipping methods.

Now, for Santa's next item on his stats wishlist. He wants to know if there's a difference between the average weights between presents under priority and economy shipping. It could give us a little more intuition to work off of when we load up the sleigh. We're very much a data-driven business, you see!

You asked for the means of the two subsamples, which can also be considered as random samples. So, here they are!

```
## # A tibble: 2 x 3
##   shipping     n  mean
##   <fct>    <int> <dbl>
## 1 economy      9  7.22
## 2 priority     9  6.89
```

the means are the same
$\mu_E = \mu_P$

What hypotheses can we test and how can we test them based off of these data?
We can test the null hypothesis that _____ **using a** permutation **. We are choosing this test because we don't have that large of a sample size and need to simulate the case if we did. Our observed difference in means is** -0.44 **.**

How can we interpret the observed difference? **On average, we expect** priority to weigh 0.44 lbs less than economy.

Simulating a null distribution.

Here's a preview of the data we were looking at earlier, but subsetted to just see present weights and shipping methods.

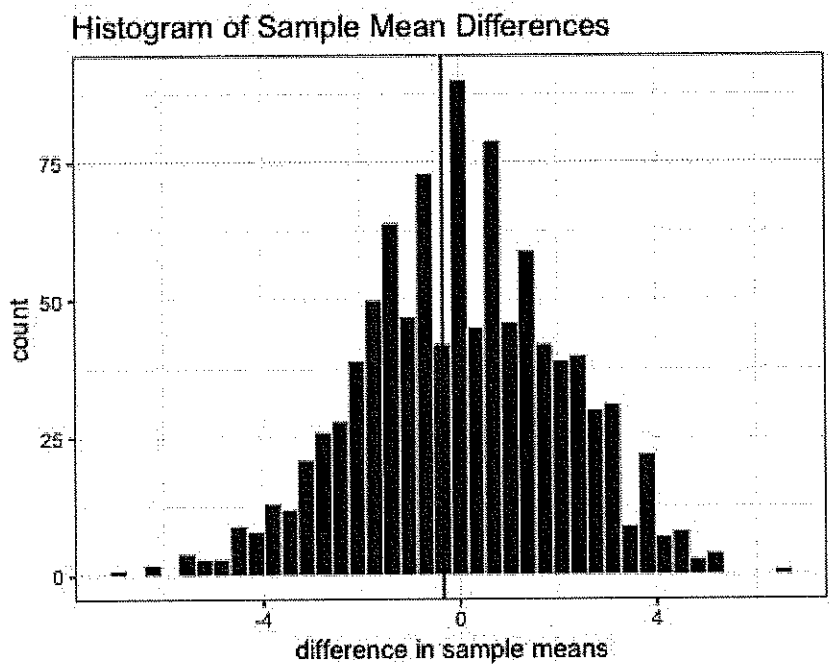
```
##   present_weights shipping
## 1               5 priority
## 2               9  economy
## 3               6 priority
## 4               8  economy
## 5               3 priority
## 6               2  economy
```

And here's a preview (the first five columns only) of what the computer is giving us after running the code you wrote to shuffle the shipping labels.

```
##   present_weights destinations shipping_i1 shipping_i2 shipping_i3
## 1               5 north america    priority    priority     economy
## 2               9 south america     economy     economy    priority
## 3               6         africa    priority    priority    priority
## 4               8           asia     economy     economy    priority
## 5               3         europe    priority    priority     economy
## 6               2      australia     economy     economy    priority
```

I understood how you shulffed the labels, but I actually don't understand how you got to this next part with all these sample means. What exactly did you do and why?

```
##   priority  economy mean_difference
## 1 6.888889 7.222222      -0.3333333
## 2 7.000000 7.100000      -0.1000000
## 3 6.454545 8.000000      -1.5454545
## 4 8.166667 6.500000       1.6666667
## 5 6.916667 7.333333      -0.4166667
## 6 8.166667 4.833333       3.3333333
```



We calculated the mean difference **for each of the new relabels so that we could simulate the** null **distribution. From this distribution, we can now calculate the** p-value **which is the probability of observing our data or more extreme, given that** $\mu_E = \mu_P$ **!**

Therefore, we can calculate our p-value ~~by seeing how much data is above our original difference.~~

? The p-value is not always probability above ...

Here is the output from our calculation:

```
## prob_below_value prob_above_value
##            0.432            0.568
```

In this case, we have exact probabilities of how many counts are in each bar.

The test will be more clear!

I would theoretically calculate p-value by taking the probability below $Diff_{OBS} = -0.44$, then get the $P(X > 0.44)$ for this distribution since its center is 0. I would add those 2 probabilities.

at least 0.432, BIG, approximately 2 * 0.432

Because our p-value is ↑ **, we** fail to reject **our null hypothesis.**
Therefore, when you pack up your next sleigh
keep doing your thang, Santa!!!

Well, that's all great to know now! Santa will be pleased. Thanks for all your help, Cheery statistician! All of the North Pole thanks you for all your data skills. I have very few things on my wishlist this year... but I do truly wish that you have a super jolly winter break!