PM592: Regression Analysis for Data Science	Name:
HW1	Flemming Wu
Distributions, EDA, Statistical Tests, Sampling Distribution	

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- There are 4 questions and 30 points possible.

Question 1 [6 points]

Solve these problems using the probability functions in R. Create 8 separate variables that contains the answer to each, and store the variables in a tibble. Attach the code you used to create the tibble and the tibble output. Assume Z~N(0,1).

1a. [1 point]. P(Z ≤ ?) = 0.01

1b. [1 point] P(|Z| ≥ 1.96)

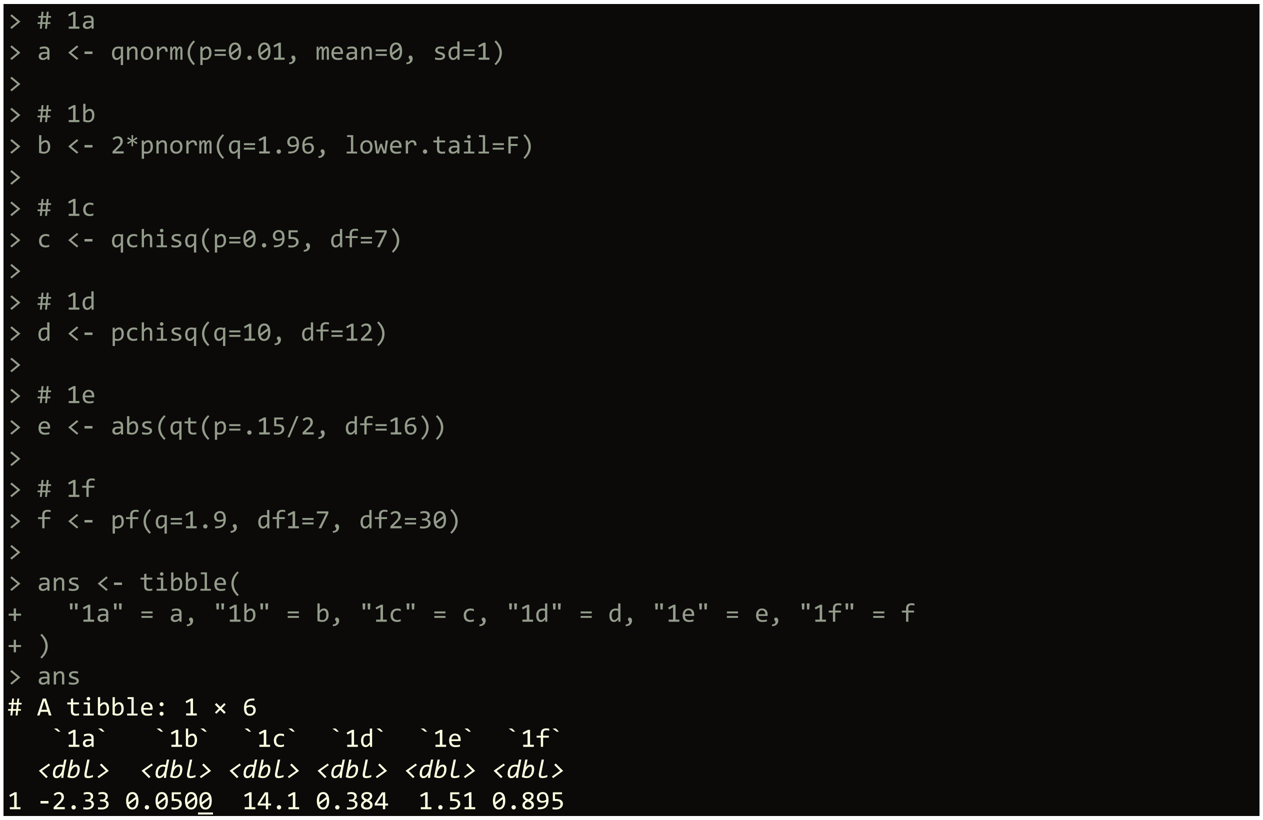
1c. [1 point] P(χ^2_7 ≤ ?) = 0.95

1d. [1 point] P(χ^2_{12} ≤ 10) = ?

1e. [1 point] P(|T_{16}| ≤ ?) = 0.15

```
abs(qt(p=(1-.15)/2, df=16)) = 0.192
```

1f. [1 point] P($F_{7,30}$ ≤ 1.9) = ?



```
> # 1a
> a <- qnorm(p=0.01, mean=0, sd=1)
>
> # 1b
> b <- 2*pnorm(q=1.96, lower.tail=F)
>
> # 1c
> c <- qchisq(p=0.95, df=7)
>
> # 1d
> d <- pchisq(q=10, df=12)
>
> # 1e
> e <- abs(qt(p=.15/2, df=16))
>
> # 1f
> f <- pf(q=1.9, df1=7, df2=30)
>
> ans <- tibble(
+   "1a" = a, "1b" = b, "1c" = c, "1d" = d, "1e" = e, "1f" = f
+ )
> ans
# A tibble: 1 × 6
   `1a`   `1b`  `1c`  `1d`  `1e`  `1f`
  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 -2.33 0.0500  14.1 0.384  1.51 0.895
```

Question 2 [9 points]

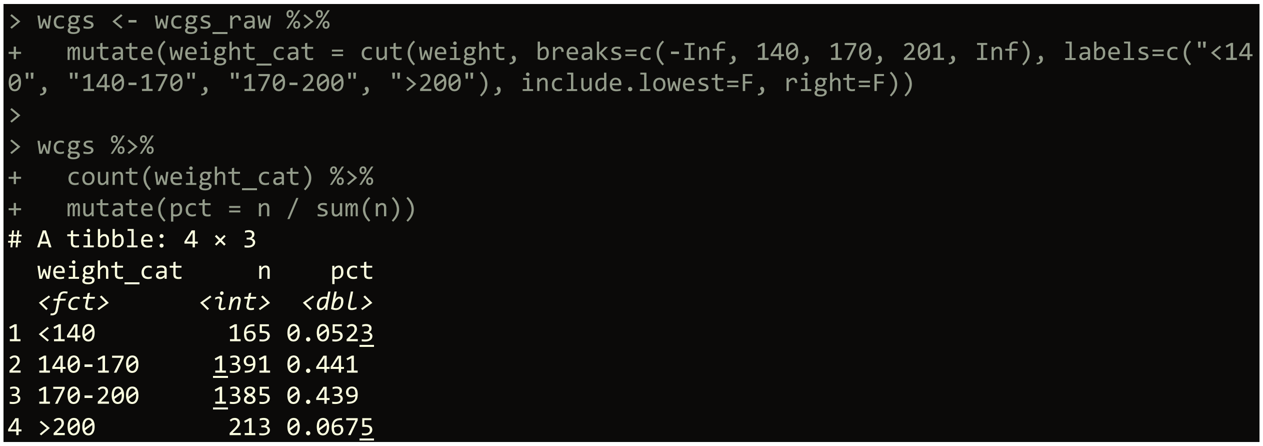
Use the “wcgs.csv” data set.

- Load the data into an object called “wcgs_raw”.
- Create a new data set called “wcgs” for the following modifications. When you are done with the problem, save the data as an R data set.

2a. [1 point] Explore this data set. What information d the “str”, “names”, and “dim” functions provide you?

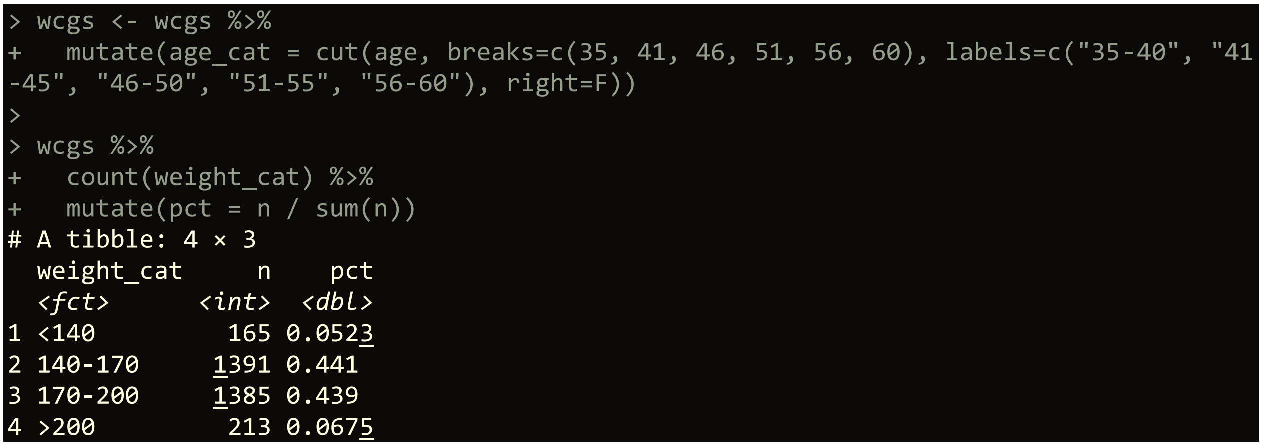
`str` provides the shape of the data frame, each column’s data type, and first few values of each column.
`names` provides a character vector containing the column names.
`dim` provides a vector of length 2, the first element being the number of rows, and the second being the number of columns.

2b. [1 point] Create an ordinal factor variable for weight category that consists of four categories: <140, 140-170, 170-200, >200. Find the frequency of each of these values in the sample.



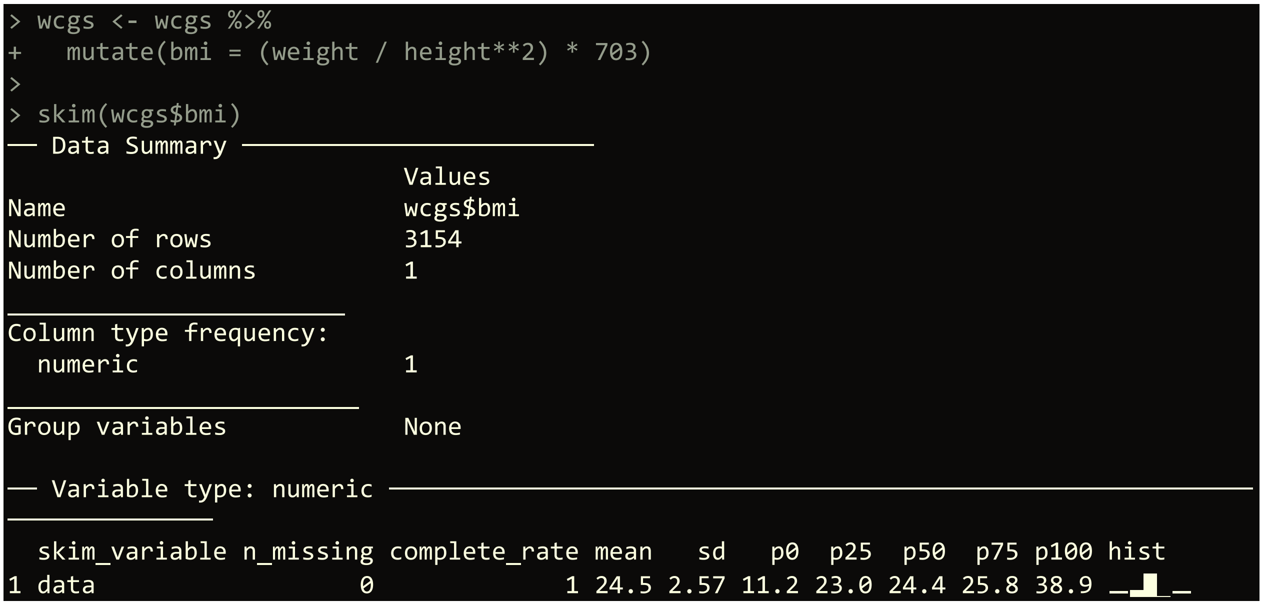
```
> wcgs <- wcgs_raw %>%
+   mutate(weight_cat = cut(weight, breaks=c(-Inf, 140, 170, 201, Inf), labels=c("<14
0", "140-170", "170-200", ">200"), include.lowest=F, right=F))
>
> wcgs %>%
+   count(weight_cat) %>%
+   mutate(pct = n / sum(n))
# A tibble: 4 × 3
  weight_cat     n    pct
  <fct>      <int>  <dbl>
1 <140         165 0.0523
2 140-170     1391 0.441
3 170-200     1385 0.439
4 >200         213 0.0675
```

2c. [1 point] Create an ordinal factor variable for age category that consists of five categories: 35-40, 41-45, 46-50, 51-55, 56-60. Find the frequency of each of these values in the sample.



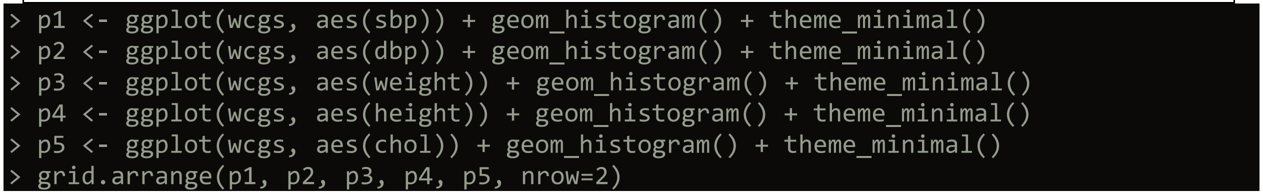
```
> wcgs <- wcgs %>%
+   mutate(age_cat = cut(age, breaks=c(35, 41, 46, 51, 56, 60), labels=c("35-40", "41
-45", "46-50", "51-55", "56-60"), right=F))
>
> wcgs %>%
+   count(weight_cat) %>%
+   mutate(pct = n / sum(n))
# A tibble: 4 × 3
  weight_cat     n    pct
  <fct>      <int>  <dbl>
1 <140         165 0.0523
2 140-170     1391 0.441
3 170-200     1385 0.439
4 >200         213 0.0675
```

2d. [1 point] Create a variable for BMI (you will have to look up the equation). In the WCGS data, height is measured in inches and weight is measured in pounds. Provide summary statistics for BMI using the package of your choice.

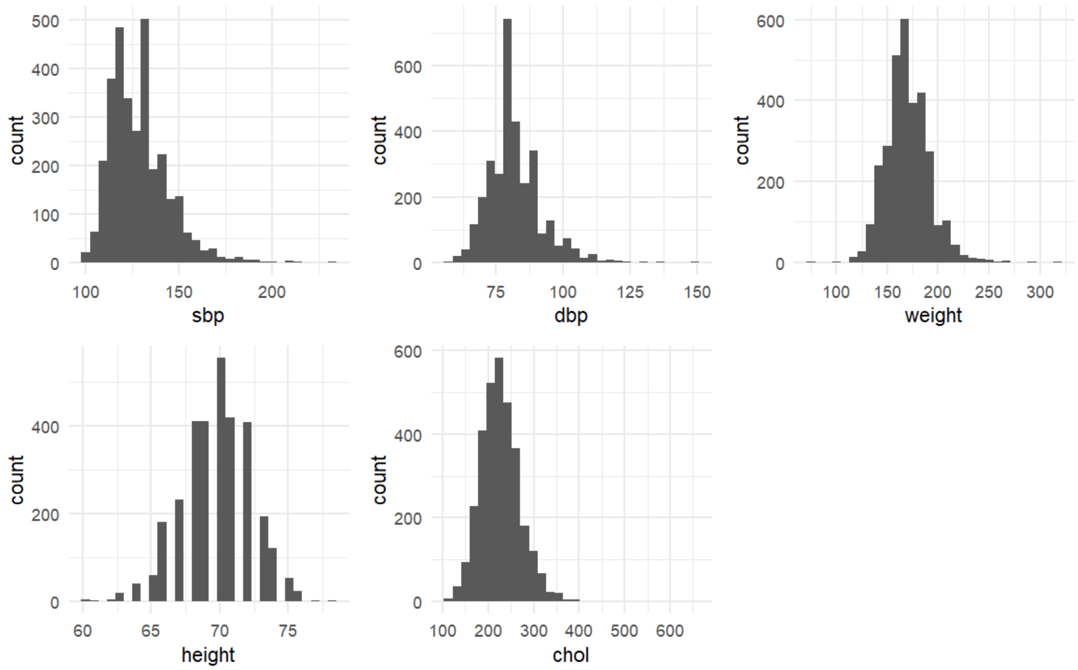


```
> wcgs <- wcgs %>%
+   mutate(bmi = (weight / height**2) * 703)
>
> skim(wcgs$bmi)
── Data Summary ────────────────────────
                           Values
Name                       wcgs$bmi
Number of rows             3154
Number of columns          1
_______________________
Column type frequency:
  numeric                  1
________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────
──────────────
  skim_variable n_missing complete_rate mean   sd   p0  p25  p50  p75 p100 hist
1 data                  0             1 24.5 2.57 11.2 23.0 24.4 25.8 38.9 ▁▃█▁▁
```

2e. [1 point] Provide histograms for age, SBP (systolic blood pressure), DBP (diastolic blood pressure), weight, height, and cholesterol. Comment on whether these variables appear normally distributed.

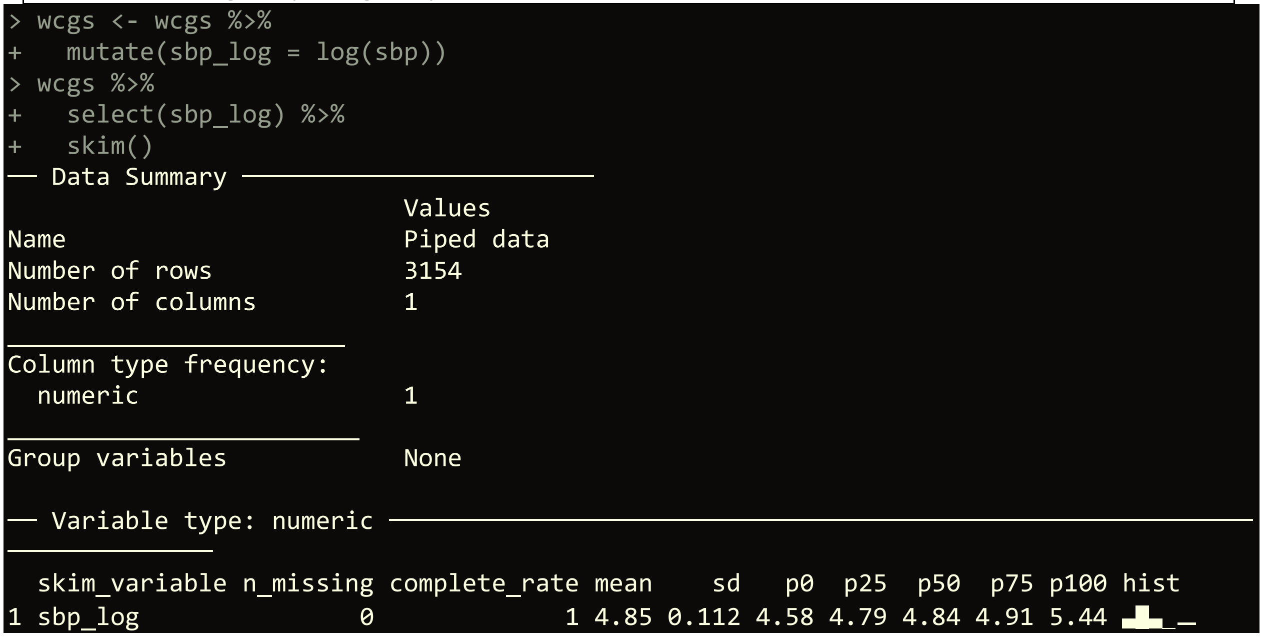


```
> p1 <- ggplot(wcgs, aes(sbp)) + geom_histogram() + theme_minimal()
> p2 <- ggplot(wcgs, aes(dbp)) + geom_histogram() + theme_minimal()
> p3 <- ggplot(wcgs, aes(weight)) + geom_histogram() + theme_minimal()
> p4 <- ggplot(wcgs, aes(height)) + geom_histogram() + theme_minimal()
> p5 <- ggplot(wcgs, aes(chol)) + geom_histogram() + theme_minimal()
> grid.arrange(p1, p2, p3, p4, p5, nrow=2)
```



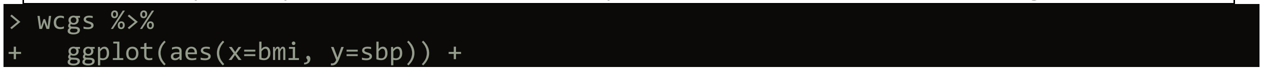
The variables: cholesterol, weight, and height appear to be about normally distributed. Diastolic blood pressure and systolic blood pressure look slightly skewed right (SBP more so than DBP).

2f. [1 point] Create a variable that is the natural log of SBP. Provide summary statistics for this variable using the package of your choice.

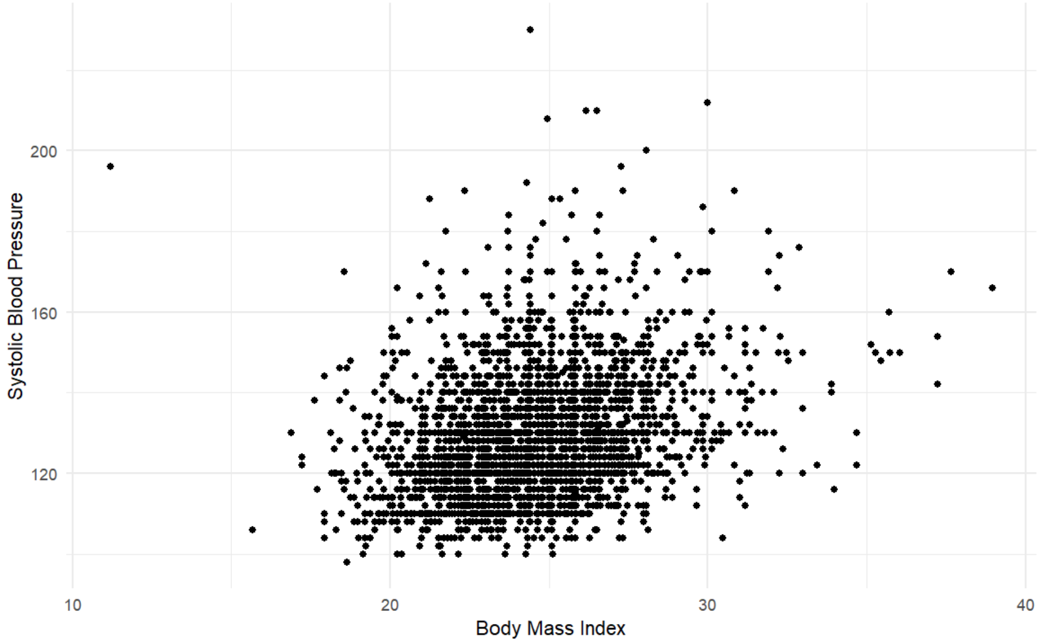


```
> wcgs <- wcgs %>%
+   mutate(sbp_log = log(sbp))
> wcgs %>%
+   select(sbp_log) %>%
+   skim()
── Data Summary ──────────────────────────
                           Values
Name                       Piped data
Number of rows             3154
Number of columns          1

_______________________
Column type frequency:
  numeric                  1

________________________
Group variables            None

── Variable type: numeric ─────────────────────────────────────────────────────────────
──────────────
  skim_variable n_missing complete_rate mean    sd   p0   p25  p50  p75  p100 hist
1 sbp_log               0             1 4.85 0.112 4.58 4.79 4.84 4.91 5.44 ▂▇▃▁▁
```

2g. [1 point] Create a scatterplot of SBP as a function of BMI. Label your axes appropriately. What are your impressions of the relationship between SBP and BMI from this figure?



```
> wcgs %>%
+   ggplot(aes(x=bmi, y=sbp)) +
+   geom_point() +
+   labs(x="Body Mass Index", y="Systolic Blood Pressure") +
+   theme_minimal()
```



From the scatterplot of BMI vs SBP, a very slight positive correlation can be seen between the two variables. However, the correlation looks very weak since the points form a blob in the middle of the plot for the most part.

2h. [1 point] Calculate the total number of cigarettes that are smoked per day by all subjects in the data set (i.e., the sum of ncigs).

```
> sum(wcgs$ncigs)
[1] 36588
```

2i. [1 point] Provide a cross-tabulation of personality type (dibpat) by smoking status. What percent of Type A personalities smoke? What percent of Type B personalities smoke?

```
> wcgs %>%
+   count(dibpat, smoke) %>%
+   group_by(dibpat) %>%
+   mutate(dibpat_count = sum(n)) %>%
+   mutate(dibpat_smoke_pct = n / dibpat_count)
# A tibble: 4 × 5
# Groups:   dibpat [2]
  dibpat smoke     n dibpat_count dibpat_smoke_pct
  <chr>  <chr> <int>        <int>            <dbl>
1 Type A No      784         1589            0.493
2 Type A Yes     805         1589            0.507
3 Type B No      868         1565            0.555
4 Type B Yes     697         1565            0.445
```

50.7% of Type A personalities smoke, and 44.5% of Type B personalities smoke

Question 3 [7 points]

Continue to use your WCGS file.

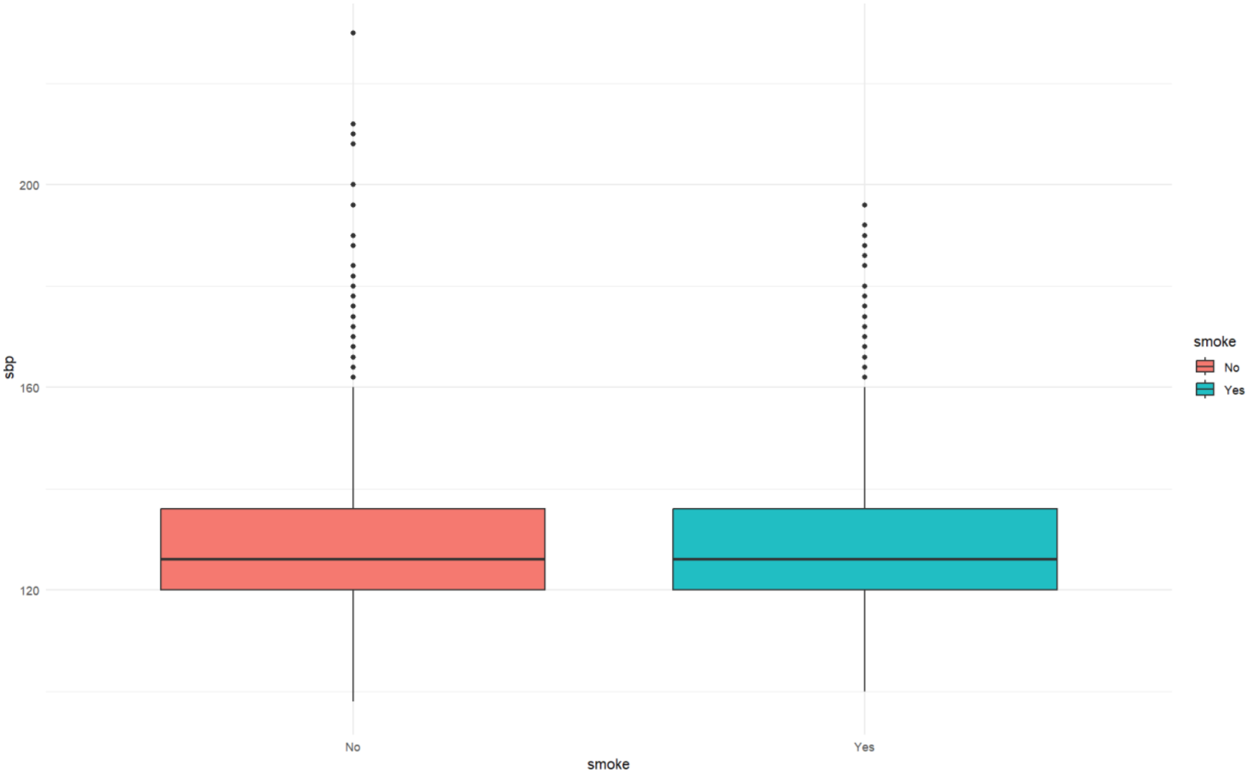
3a. [1 point] Consider a SBP of 125 as “normal”. Is there any evidence that the mean SBP of individuals in this sample is different from 125?

```
> t.test(x=wcgs$sbp, mu=125, alternative="t")

	One Sample t-test

data:  wcgs$sbp
t = 13.496, df = 3153, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 125
95 percent confidence interval:
 128.1050 129.1606
sample estimates:
mean of x 
 128.6328 
```

Since only the standard deviation of the sample is known, I opted for a one-sample, two-sided t-test to test for difference in the “normal” mean of 125 and the mean of my data set. The test statistic for the t-test was 13.496, and the p-value was 2.2 x 10^-16. Since the p-value is << 0.01, there is evidence to reject the null hypothesis that the mean SBP of individuals in the sample is not different from 125, and accept the alternative hypothesis that the mean SBP of individuals in the sample is different from 125.

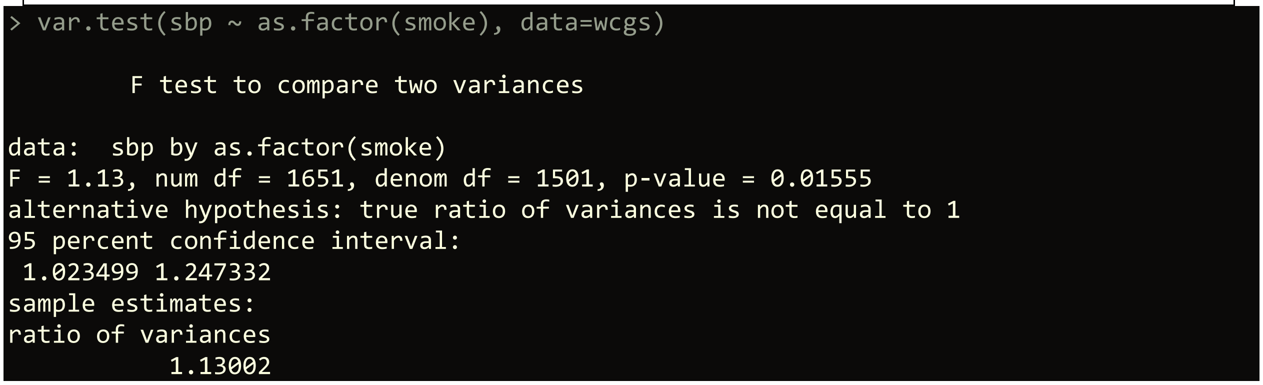
3b. [1 point] Create boxplots of SBP by smoking status. What is your impression of how SBP relates to smoking status?

```
> wcgs %>%
+   ggplot(aes(x=smoke, y=sbp, fill=smoke)) +
+   geom_boxplot() +
+   theme_minimal()
```



It appears that SBP does not differ by smoking status, the distributions of SBP seems to be very similar between the smoking group and non-smoking group.

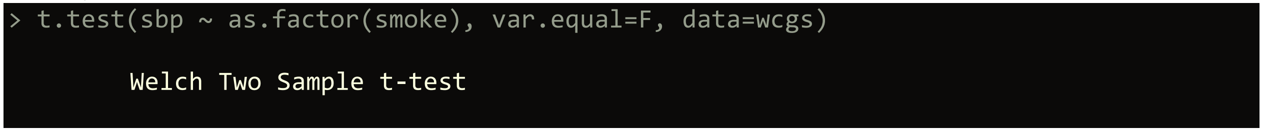
3c. [1 point] Perform a parametric statistical test to determine if mean SBP differs by smoking status.



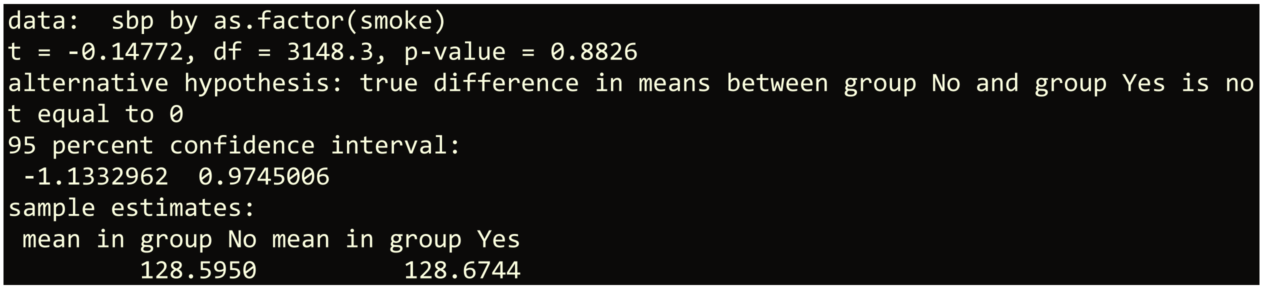
```
> var.test(sbp ~ as.factor(smoke), data=wcgs)

        F test to compare two variances

data:  sbp by as.factor(smoke)
F = 1.13, num df = 1651, denom df = 1501, p-value = 0.01555
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.023499 1.247332
sample estimates:
ratio of variances
           1.13002
```

Using a p-value threshold of 0.05, can reject the null hypothesis that the variances are equal and accept alternative the alternative that the variances are not equal.

Then, perform t-test:



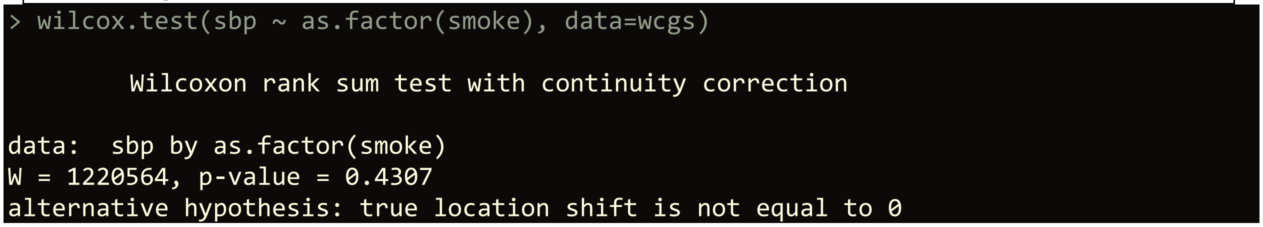
```
> t.test(sbp ~ as.factor(smoke), var.equal=F, data=wcgs)

        Welch Two Sample t-test
```



```
data:  sbp by as.factor(smoke)
t = -0.14772, df = 3148.3, p-value = 0.8826
alternative hypothesis: true difference in means between group No and group Yes is no
t equal to 0
95 percent confidence interval:
 -1.1332962  0.9745006
sample estimates:
 mean in group No mean in group Yes 
         128.5950          128.6744 
```

With the large p-value, there is no evidence to reject the null hypothesis. The null states that there is no difference in SBP between the smoking group and non-smoking group.

3d. [1 point] Perform a non-parametric statistical test to determine if mean SBP differs by smoking status.



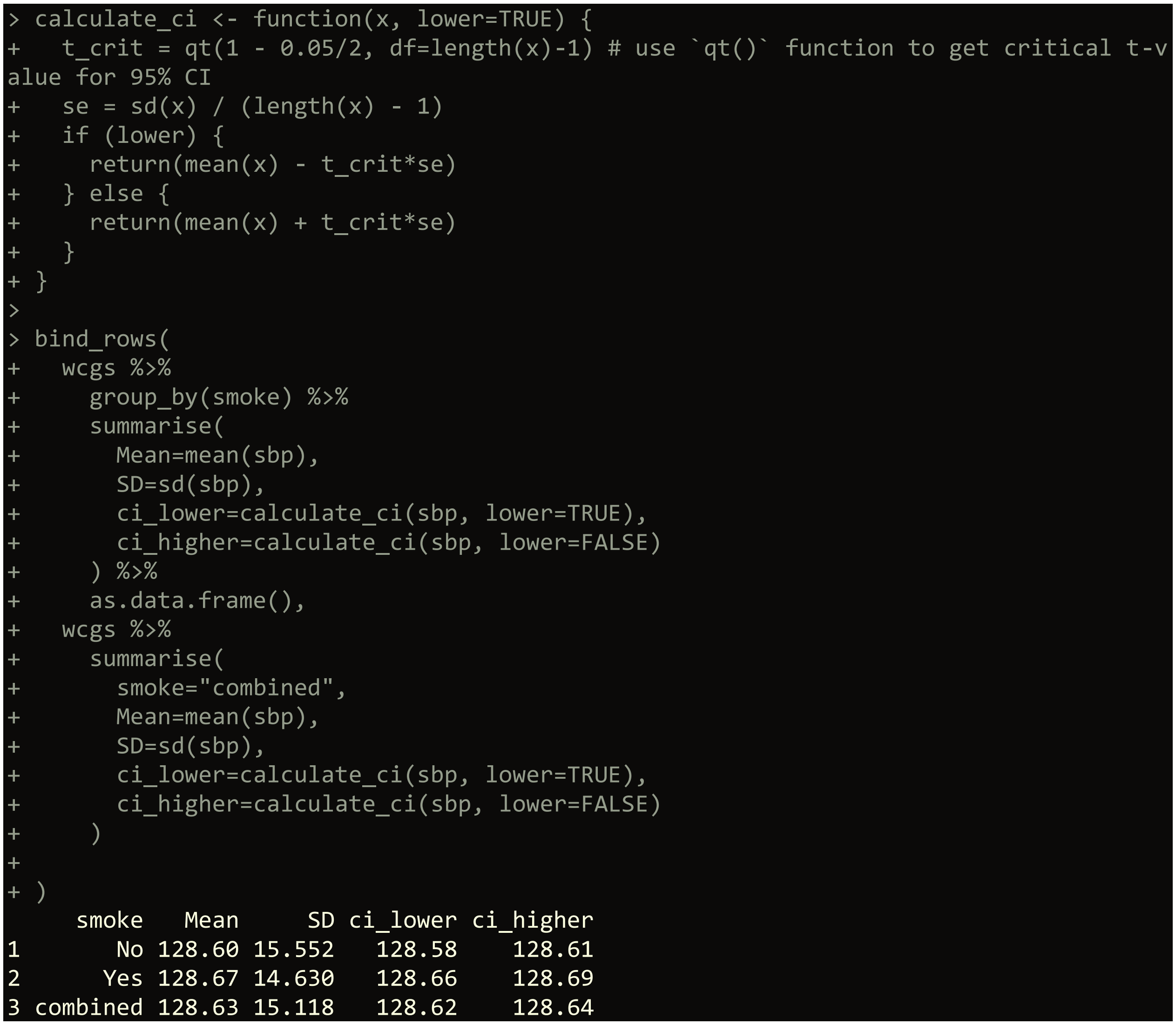
```
> wilcox.test(sbp ~ as.factor(smoke), data=wcgs)

        Wilcoxon rank sum test with continuity correction

data:  sbp by as.factor(smoke)
W = 1220564, p-value = 0.4307
alternative hypothesis: true location shift is not equal to 0
```

Again, the large p-value indicates that there is no evidence to reject the null hypothesis. The null states that there is no difference in SBP between the smoking group and non-smoking group.

3e. [3 points] Provide a conclusion about the hypothesis that there is a difference in mean SBP between smokers and nonsmokers. Provide a brief written summary of your methods and results. Make sure to justify your choice of statistical test. Include a small table that shows descriptive statistics (mean and 95% CI) for SBP overall and for SBP stratified by smokers and non-smokers. Include the p-value for comparing the smoking groups.

In order to test whether there was a difference in mean SBP between smokers and non-smokers, a statistical test can be used. Two common tests for testing the difference in means between two groups are the t-test and the Wilcoxon Rank-Sum Test, with the former being parametric and the latter non-parametric. The parametric t-test assumes that observations in each group are independent, which I determined is safe to assume, as SBP level in one group is not influenced by or related to SBP levels in the other group. It also assumes normality, which can be safely assumed given the Central Limit Theorem, which states that for large sample sizes, the sampling *distribution* of the mean will approach a normal distribution regardless of the distribution of the population mean. Since the sample size is above 3,000 a normal distribution of means can be assumed. Finally, a variance test was conducted to determine which t-test to use, as the standard t-test assumes equal variances between the two groups. A p-value of 0.05 was used to determine the threshold of the variance test, and since the observed p-value was smaller than this value, the alternative hypothesis that the variances between the two groups was accepted. Finally, a Welch's t-test, a t-test that is used when the assumption of equal variances is not met was used to test if SBP varied by smoking status. The p-value from the test was **0.8826**, which was not small enough to reject the null hypothesis that SBP did not vary between smokers and non-smokers.

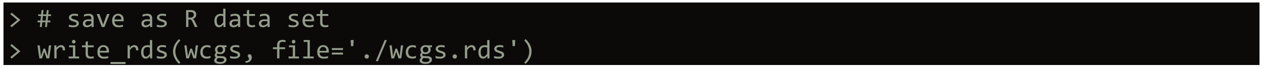
Below is the code used to generate a small summary statistics table comparing means, standard deviation, and 95% CI of the mean for smokers, nonsmokers, and combined.



```
> calculate_ci <- function(x, lower=TRUE) {
+   t_crit = qt(1 - 0.05/2, df=length(x)-1) # use `qt()` function to get critical t-v
alue for 95% CI
+   se = sd(x) / (length(x) - 1)
+   if (lower) {
+     return(mean(x) - t_crit*se)
+   } else {
+     return(mean(x) + t_crit*se)
+   }
+ }
>
> bind_rows(
+   wcgs %>%
+     group_by(smoke) %>%
+     summarise(
+       Mean=mean(sbp),
+       SD=sd(sbp),
+       ci_lower=calculate_ci(sbp, lower=TRUE),
+       ci_higher=calculate_ci(sbp, lower=FALSE)
+     ) %>%
+     as.data.frame(),
+   wcgs %>%
+     summarise(
+       smoke="combined",
+       Mean=mean(sbp),
+       SD=sd(sbp),
+       ci_lower=calculate_ci(sbp, lower=TRUE),
+       ci_higher=calculate_ci(sbp, lower=FALSE)
+     )
+
+ )
     smoke   Mean     SD ci_lower ci_higher
1       No 128.60 15.552   128.58    128.61
2      Yes 128.67 14.630   128.66    128.69
3 combined 128.63 15.118   128.62    128.64
```

As can be seen in the table, the SBP smeans for smokers and non-smokers are almost identical with non-smokers having a mean of 128.60, smokers having a mean of 128.67, and combined the mean is 128.63.

Saving RDS file:



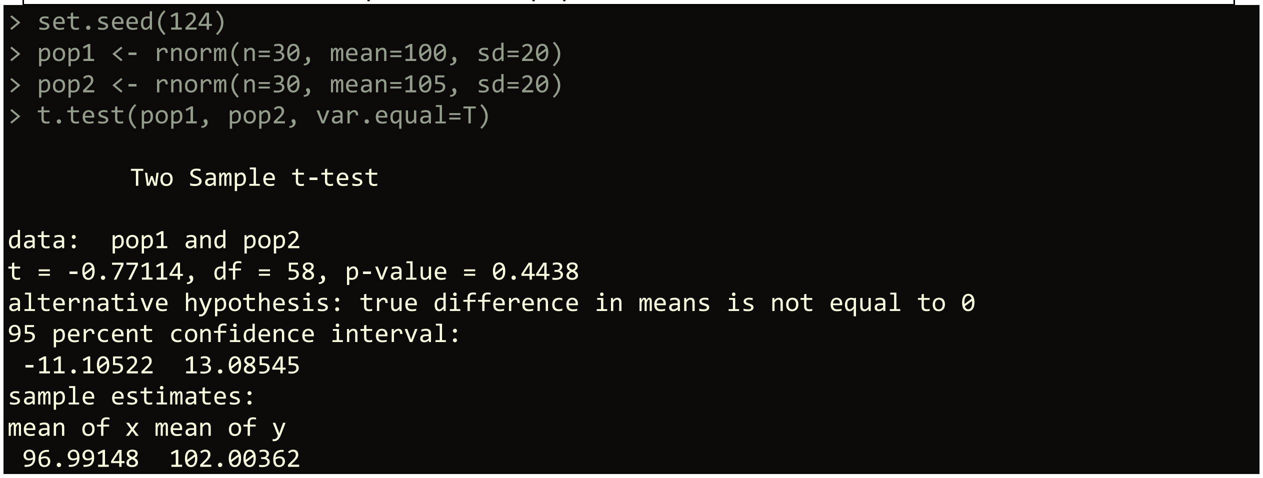
```
> # save as R data set
> write_rds(wcgs, file='./wcgs.rds')
```

Question 4 [8 points]

In this problem we will simulate drawing samples from two different populations.

Use the “help” (?) function to determine how the “rnorm” function works.

4a. [2 points] Use the “rnorm” function to create a vector named “pop1” that contains of 30 samples from a population with $\mu = 100, \sigma = 20$. Create a vector named “pop2” that contains 30 samples from a population with $\mu = 105, \sigma = 20$. Would the results of a t-test lead us to conclude that these samples are from populations with different means?



```
> set.seed(124)
> pop1 <- rnorm(n=30, mean=100, sd=20)
> pop2 <- rnorm(n=30, mean=105, sd=20)
> t.test(pop1, pop2, var.equal=T)

        Two Sample t-test

data:  pop1 and pop2
t = -0.77114, df = 58, p-value = 0.4438
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.10522  13.08545
sample estimates:
mean of x mean of y
 96.99148  102.00362
```

The high p-value of 0.44 signifies that there is not enough evidence to reject the null hypothesis, which states that the samples are from populations with different means. I suspect that this is due to the smaller sample size.

4b. [4 points] Re-run the program several (>6) times, each time altering the value of the means, standard deviations, OR sample sizes (only alter one parameter at a time). Fill the values you obtain into the table below.

	Population 1		Population 2		Sample 1			Sample 2				
Model	μ_1	σ_1	μ_2	σ_2	N	$\bar{x}$	SD	N	$\bar{x}$	SD	t	p
1	100	20	105	20	30	98.99	17.25	30	102.00	12.66	-0.18	0.44
2	100	20	105	20	300	99.64	19.89	300	103.38	19.09	-2.35	0.019
3	100	20	105	20	100	100.23	20.88	100	102.17	20.39	-0.67	0.51
4	100	50	105	50	30	87.26	56.84	30	104.42	52.13	-1.22	0.23
5	100	5	105	5	30	99.18	4.78	30	105.45	4.71	-5.12	3.69e-06
6	90	20	110	20	30	83.72	19.34	30	119.80	24.31	-6.36	3.42e-08
7	95	20	105	20	30	94.98	24.33	30	98.71	20.02	-0.65	0.52
8	100	75	100	75	30	85.55	65.27	30	90.58	74.93	-0.28	0.78

4c. [2 points] Summarize the effect of changing the population mean, standard deviation, and N has on the t and p-values. What general trends do you observe?

Given that the population means are different, it seems that increasing the sample size increases the power of the t-test to determine that the samples come from means from different populations, and lowering the sample size decreases the t-test’s power. As for standard deviation, it seems that with lower standard deviations, the ability for a t-test to detect difference in means from two populations where the difference in means and sample sizes are both small is better. Increasing the standard deviation seems to have the opposite effect, where it is harder to detect differences in means. Lastly, increasing the difference between population means seems to increase the power of the t-test, given smaller sample size.