

## FINAL

1. Let the null hypothesis be that there is no difference in death rates for men and women, while the alternative hypothesis state that the death rates between men and women differ significantly. Because the data is in the form of a ratio (death rate), and there are two levels (one for men and another for women), we must oconduct a two-way ANOVA in order to determine the existence of a signifcant difference. The two-way ANOVA found the p-value to be 0.0112. Since it is not less than 0.05, we fail to reject the null hypothesis that there is a significant diffrence in the number of deaths between men and women.
2. The racial group with the highest death rate is American Indian, with 0.0714. Despite the wide margin between the death rate in American Indian populations and other populations, the resulting anova did not identify a significant difference between races.

```
52
53
54 |
55
56 death_rates <- data[, c("Race", "Death_yn")]
57 death_rates$Death_yn <- as.numeric(lapply(death_rates$Death_yn, mod2))
58 head(death_rates); shapiro.test(death_rates$Death_yn); print("passes shapiro wilks")
59
60 #leveneTest(Death_yn ~ Race, data=death_rates)
61
62 chisq.test(death_rates$Death_yn)
63 #print("Since the p-value is greater than 0.05, the test shows that
64 #   race does not   significant effect on death rate(definitely does, but why doesn't this show it??)")
54:1 (Top Level) ↕
```

Console

Terminal ×

Jobs ×

~/Documents/school/MATHSTATS/final/ ↗

[1] "passes shapiro wilks"

```
> #leveneTest(Death_yn ~ Race, data=death_rates)
```

```
>
```

```
> chisq.test(death_rates$Death_yn)
```

Chi-squared test for given probabilities

data: death\_rates\$Death\_yn

X-squared = 3423, df = 3599, p-value = 0.9822

```
> #print("Since the p-value is greater than 0.05, the test shows that
```

```
> #   race does not   significant effect on death rate(definitely does, but why ..." ... [TRUNCATED]
```

Warning message:

```
In chisq.test(death_rates$Death_yn) :
```

Chi-squared approximation may be incorrect

```
> |
```

- The population of age 80 or older showed the greatest death rate of the other age groups. After conducting a one-way ANOVA, the resulting p-value was less than  $2e-16$ , which is less than 0.05. So, we can reject the null that there is no significant difference between those older than 80 and other age groups.

```

1 data <- read.csv('Final4.csv')
2
3 getlabel <- function(x,s) x[x$Age == s,]
4 deaths <- function(x) as.numeric(lapply(x$Death_yn, function(s) if(s=="Yes") 1 else 0))
5 death_rate <- function(L) sum(L[L == 1])/length(L)
6 process <- function(x,s) death_rate(deaths(getlabel(x,s)))
7
8 d3039 <- process(data, "30 - 39 Years")
9 d1019 <- process(data, "10 - 19 Years")
10 d4049 <- process(data, "40 - 49 Years")
11 d7079 <- process(data, "50 - 59 Years")
12 d2029 <- process(data, "20 - 29 Years")
13 d09 <- process(data, "0 - 9 Years")
14 d6069 <- process(data, "60 - 69 Years")
15 d80p <- process(data, "80+ Years")
16
17 max(d3039, d1019, d4049, d7079, d2029, d09, d6069, d80p)
18 print('The age group with teh highest death rate is 80+')
19
20 death_rates <- data[, c("Age", "Death_yn")]
21 death_rates$Death_yn <- as.numeric(lapply(death_rates$Death_yn, mod2))
22 model <- aov(Death_yn ~ Age, data=death_rates)
23
24 summary(model)

```

28:1 (Top Level) ⚡

Console

Terminal ×

Jobs ×

~/Documents/school/MATHSTATS/final/ ↗

```
> model <- aov(Death_yn ~ Age, data=death_rates)
```

```
> summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	8	39.74	4.967	138.7	<2e-16 ***
Residuals	3591	128.56	0.036		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

- To determine if there is an association between two variables, a Chi-Squared Test of Independence must be conducted. Let the null hypothesis be the claim that seriology level and death rate are independent of each other and the alternative hypothesis that they are associated. After running the ANOVA, the p-value was found to be less than 0.05. So, the null hypothesis is rejected and it can be concluded that there is an association between seriology and death rate. This differs with respect to gender. For if, without loss of generality, the same null and alternative hypotheses are made, the ANOVA reports a p-value greater than 0.05. So, there is no association between gender and death rate.

5. Let the null hypothesis claim that there is no association between cholesterol and seriology and the alternative hypothesis claim that there is. Since this implicitly assumes independence, we can proceed with running a Chi-Squared Test of Independence. The test reports a p-value less than  $2e-16$ . So, the null hypothesis is rejected. Therefore, there is an association to be made between cholesterol and death rate.

```
1 data <- read.csv('Final4.csv')
2 ser <- data[, c("Cholesterol", "Ser")]
3 nser <- ser[ser$Ser>0,]
4 str(nser)
5
6 model <- aov(Ser ~ Cholesterol, data=nser)
7 summary(model)
8 print("Since the p-value is greater than 0.05, we
9       accept the null hypothesis that there is
10      an association between cholesterol and seriability")
11
```

10:59 (Top Level) ↕

**Console**

**Terminal** ×

**Jobs** ×

~/Documents/school/MATHSTATS/final/ ↗

Pearson's Chi-squared test

data: nser

X-squared = 13606, df = 3597, p-value < 2.2e-16

```
> model <- aov(Ser ~ Cholesterol, data=nser)
```

```
> summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cholesterol	1	15	15.25	0.161	0.689
Residuals	3596	341538	94.98		