Yellow Fever in Mali:

Anaysis of a Single Mean

Carly Lundgreen

Introduction and Data Description

According to the CDC, the yellow fever virus, which is primarily found in sub-tropical areas of Africa and South America, can cause illness ranging from a fever with mild aches and pains to severe liver disease accompanied by bleeding and jaundice. The disease originates from the bites of infected mosquitoes, and there is currently no anti-viral drug available to treat or cure infection. Preventative treatment, such as wearing protective clothing and receiving vaccinations, are essential to avoid contracting this virus. The WHO reports that 34 African countries are endemic for yellow fever, and a modeling study based on African data sources estimated the burden of yellow fever during 2013 was 84,000-170,000 severe cases and 29,000-60,000 deaths. This virus, however, is prevented by a very effective vaccine. The WHO states that one dose of the vaccine is sufficient to grant sustained immunity and life-long protection against the disease, although some studies say that a booster may be necessary after about ten years. Regardless, vaccination is incredibly important for the people living in regions that are endemic for yellow fever.

These data were gathered through a longitudinal cohort study that was conducted to measure long-term immunity against yellow fever in children vaccinated during infancy. A group of 587 healthy infants, each around the age of 9 months, were vaccinated in the country of Mali and followed for about 4.5 years. The concentration of yellow fever antibodies in each child was measured and recorded before vaccination and at the time of follow-up. A microneutralization assay was used to measure the amount of protective antibodies in each child. The concentration of antibodies was standardized with reference to the yellow fever WHO International Standard, and is reported in units of IU/mL. The international unit, IU, is a unit of measurement for the amount of a substance that is present in a sample. The mass or volume that constitutes an IU varies depending on the substance that is being measured, which makes conversion to standard metric measurements (such as grams) difficult. For this reason, we will keep the units of antibody concentration as IU/mL in our discussion.

So, just how effective is the yellow fever vaccine? We will attempt to discover whether there is a significant presence of antibodies for the average Malian infant after the (approximate) 4.5 year period following vaccination by performing a one-sample t test for the mean concentration of antibodies in the blood.

Exploratory Data Analysis

We first note that the standardized concentration measurement for yellow fever antibodies before vaccination was zero for each of the 587 children in the study, indicating the absence of antibodies pre-vaccination for each child. This means that the difference in antibody concentration is the same as the concentrations measured after the (approximate) 4.5-year period, so from this point on I will refer to the difference in antibody concentration (after vaccination minus before vaccination) simply as "antibody concentration."

The left plot in Figure 1. displayed below, illustrates the distribution of antibody concentrations for the 587 children after the end of their follow-up period. The significant right-skewness would be a concern to us if not for the large sample size of this study, which we will discuss in the following section. The distribution of the number of days elapsed since receiving the vaccination is displayed in the rightmost plot of Figure 1. This study did not follow up with each child at the same time, and we will consider this in the discussion of our results. For reference, red lines have been placed on the histogram to denote the 4, 4.5, and 5-year post-vaccination mark.

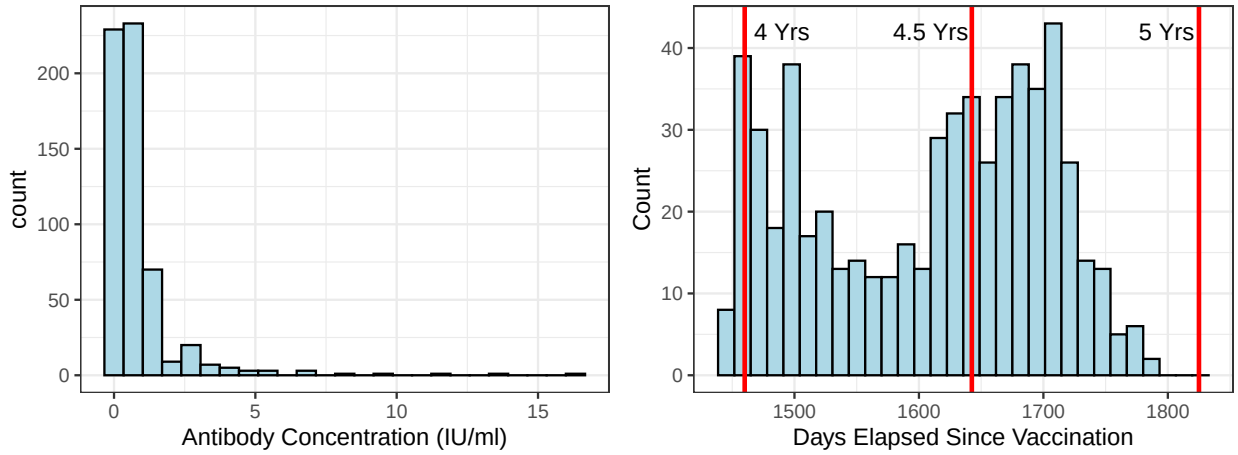


Figure 1: These plots show the distribution of antibody concentration after about 4.5 years for 587 Malian children and the distribution of the days elapsed since vaccination

Figure 2, displayed below, explores the distribution of yellow fever antibody concentrations in infants by both gender and age. In this study, infants ages 9,10,11, and 12 months received yellow fever vaccinations for for the first time. After the (approximate) 4.5 year mark, we can clearly see that the children who received vaccinations at 9 months had a much larger spread of antibody concentrations in their blood than children vaccinated at 10, 11, or 12 months. We can also see that the spread of antibody concentrations for female infants was greater than the spread of antibody concentrations for male infants. However, this difference appears to arise from only a few outliers. The IQR for both male and female infants is comparable. A similar observation can be made about the distributions of antibody concentrations by age.

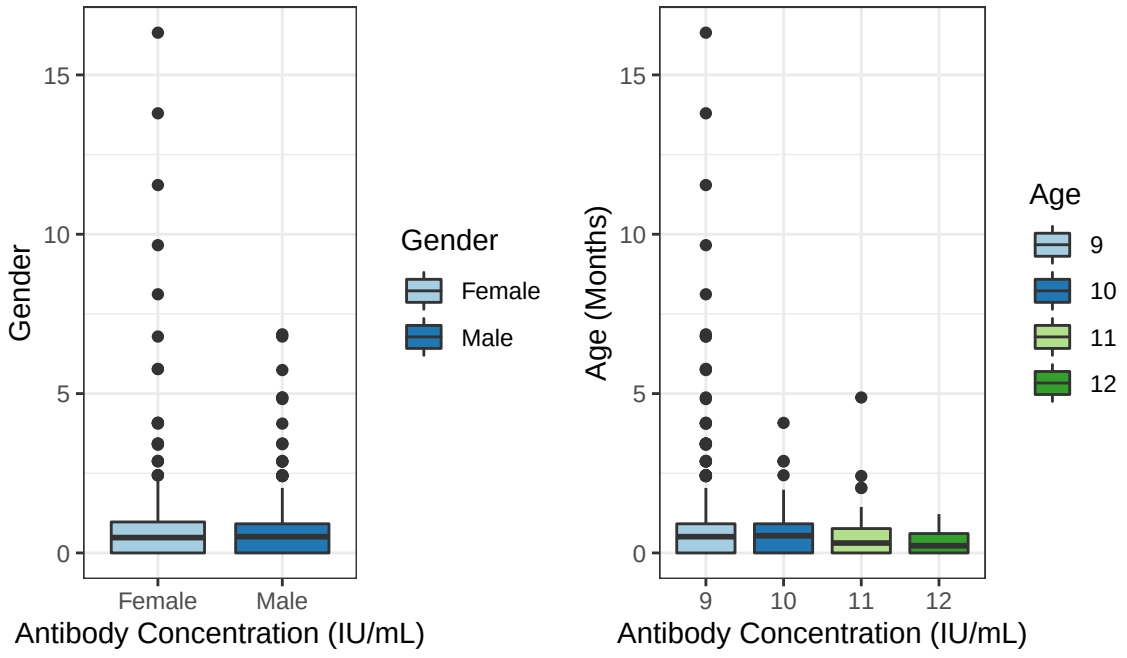


Figure 2: Two plots showing the distributions of antibody concentrations by gender and by age

Finally, below is a table of summary statistics for the distribution of antibody concentrations. Note that

Table 1: Summary statistics for yellow fever antibody concentrations for Malian infants after vaccination

Mean	SD	Median	IQR	Min	Max
0.83	1.45	0.5101568	0.9146311	0	16.32502

the minimum antibody concentration was zero, indicating that one or more children still did not have yellow fever antibodies after receving the vaccine. This raises the question-under what conditions is a greater concentration of antibodies developed in the blood? This would be a potentially interesting route for further analysis of these data. For now, we will proceed with our analysis of the mean antibody concentration after the approximate 4.5 year period.

Analysis

Before beginning this analysis, we will examine whether the assumptions of a one-sample t procedure are sufficiently met. First, we consider independence. It appears reasonable to assume that antibody detection is independent between children. However, if children from multiple families were included in this study, there may be some level of dependence between those children in response to the vaccination based on genetic factors. Though the study does not provide information about whether some of the children came from the same family, I believe that that (potential) level of dependence is likely too small to matter in this scenario. Secondly, we will consider the condition of normality. Because our sample data is so clearly right-skewed, we rely on the Central Limit Theorem. Our sample size of 587 means that the sampling distribution of sample means will be sufficiently normal for us to utilize one-sample t procedures. The assumptions for a one-sample t procedure appear to be met, so we will proceed with the analysis.

A common concentration threshold to determine whether a blood sample is seropositive for yellow fever, which means it has detectable yellow fever antibodies, is 0.5 IU/mL. We will utilize this standard concentration to determine whether the mean antibody concentration for the 587 Malian children after receiving a yellow fever vaccine is greater than 0.5 IU/mL. In other words, we will analyze whether the average child has significantly detectable antibodies after being vaccinated (in this case approximately 4.5 years after being vaccinated). Our hypotheses for this test are below. Note that we will use a significance level of $\alpha = 0.5$ for our analysis.

$$H_o: \ \mu = 0.5$$
$$H_a: \ \mu > 0.5$$

The results of the test are displayed in the table below. The p-value of 5.46 is clearly less than our significance level of 0.05, so we conclude that the mean concentration of yellow fever antibodies in children, about 4.5 years after receiving the vaccination as an infant, is greater than 0.5 IU/mL, the standard threshold for yellow fever seropositivity. For reference, we also calculated a 95% confidence interval for the average concentration of yellow fever antibodies, and this was found to be (0.71,0.94), also in units of IU/mL.

Table 2: Results from One-Sample t test for the mean yellow fever antibody concentration in Malawian infants

Test-Stat	P-value
5.46	0

Conclusion

Based on the results of this test, the average yellow fever vaccination in Malian infants leads to seropositivity somewhere within approximately 4.5 years. The importance of yellow fever antibodies in sub-saharan African children cannot be overstated, as it nearly guarantees their immunity from the virus. This will lead to a significant increase in quality of life for these children and their families. We must consider, however, that some of the literature regarding yellow fever vaccinations seems to be questioning whether lifetime immunity is truly possible with a single vaccination in infancy. Some say that a booster may be needed after approximately ten years.

Because the follow-up times for this study were all less than five years, it would be interesting if antibody data could once again be gathered from these children to see whether the antibody concentrations had changed. We also note that the antibody measurements from this study were not taken in unison, but were spread out over a number of months. A potential space for further analysis could be to compare these study results to a study in which patient vaccination and follow-up took place at the same time, although in the world of clinical studies in sub-saharan Africa, this is likely not very feasible. Regardless, we can conclude that yellow fever vaccination in infancy does appear to lead to significant antibody concentration after the period of a few years.

Alcohol-Impaired Driving:

Analysis of Two Means

Carly Lundgreen

Introduction and Data Description

The CDC reports that in 2018, 10,511 people were killed in alcohol-impaired driving (AID) crashes in the United States, which accounted for 29% of all traffic-related deaths in the country that year, and based on state-level data, this ends up being an average of about 210 deaths per state that year. According to the CDC, rates of AID and alcohol-involved fatal crashes have gone down in recent years. However, alcohol-impaired drivers still got behind the wheel of vehicles about 147 million times in 2018. Needless to say, alcohol-impaired driving is a very significant problem on US roadways, with consequences reaching far beyond these fatality numbers.

We are interested in discovering whether the CDC's claim that deaths from AID has decreased significantly in recent years is supported by these data. It is important to note that an 'alcohol-impaired driver' has been defined as a driver with a blood-alcohol level that is above 0.08 g/dL. The data for this analysis was retrieved from the CDC website. The dataset provides the number of people killed in alcohol-impaired driving crashes from 2003-2012 by state. The number of deaths has been aggregated so that there is a single variable, `deaths`, for the entire 9-year period. We opted to calculate the 9-year average for each of the states and create a new variable, `avg_deaths`. We will compare the state-by-state 9-year average to the state-by-state number of deaths from AID in the year 2018, which data was also retrieved from the CDC's website.

Exploratory Data Analysis

Table 1, shown below, compares summary statistics for the 9-year average number of deaths per state due to alcohol-impaired driving (2003-2012) and the average number of deaths per state from AID in 2018. We can see that both the median and mean deaths decreased over time, which does appear to support the CDC's claim. In Figure 1, we have mapped AID deaths by state for both the 9-year average and the 2018 values. To the naked eye, these maps appear identical, indicating that the distributions of deaths likely has not changed very significantly between each time period. That is, states that had very high numbers of deaths still have high numbers of deaths, and states that had low numbers of deaths still have low numbers of deaths. Figure 2, on the following page, displays the distributions of deaths for each time period in side-by-side boxplots. Three outliers can clearly be seen for each group. These outliers for AID deaths, in both cases, are in the states of Florida, California, and Texas, respectively. These states can easily be spotted in Figure 1 with high numbers of deaths in both time periods. We will discuss these outliers further in the following section.

Table 1: Summary statistics for AID deaths by state for both 2018 and the average over the 9-year period 2003-2012

	Mean	SD	Median	IQR	Min	Max
2003-2012	250.90	269.08	179.72	249.61	22.11	1463.44
2018	210.06	259.18	135.50	183.00	15.00	1439.00

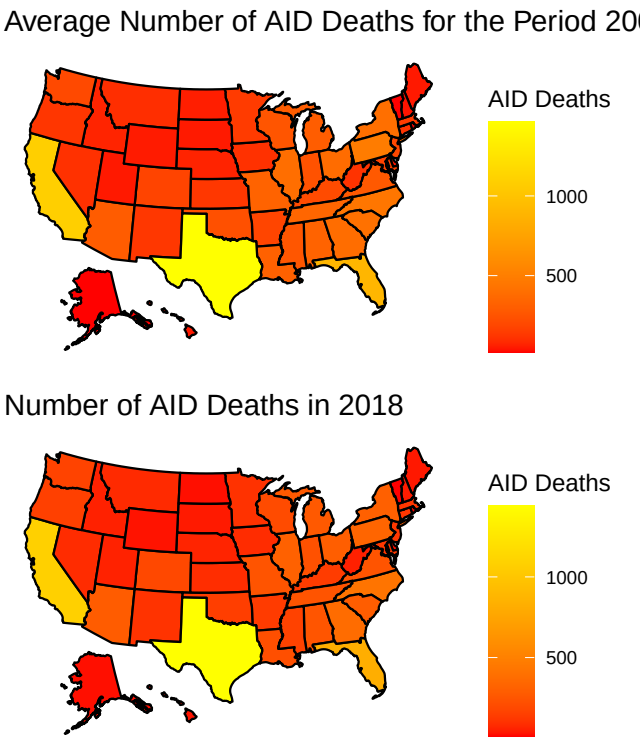


Figure 1: AID Deaths Mapped by State

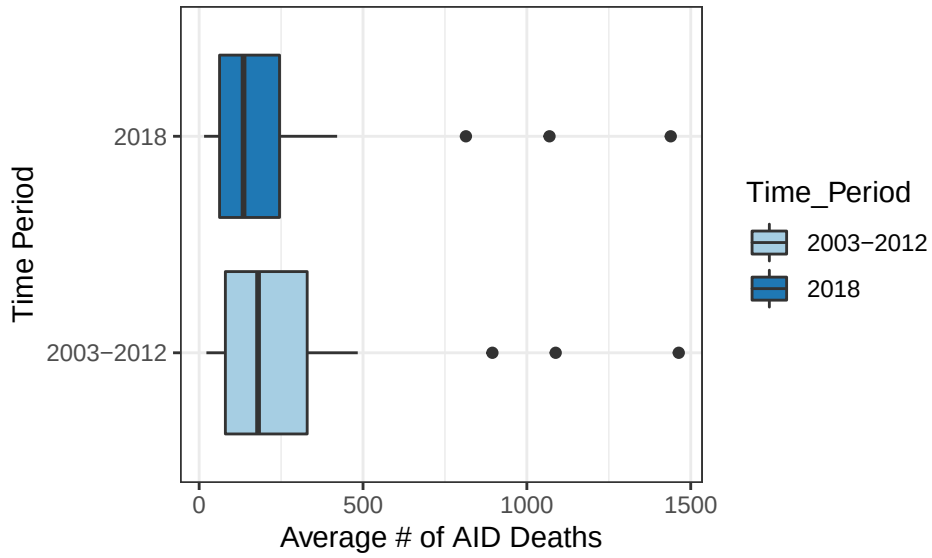


Figure 2: Distributions of the average number of AID deaths by state from 2003-2012 and the number of AID deaths by state in 2018.

Analysis

To determine whether the CDC's claim is supported by these data, we will perform a two-sample t test for a difference in means. However, we must first consider the test assumptions before proceeding. The first assumption is that our samples are distributed normally, or that we have a large enough sample size to employ the Central Limit Theorem. In this case, we saw in the side-by-side boxplots that there are outliers in each group that cause a significant right-skewness. However, the sample size for each group is 50 (one observation for each state), which is large enough to conclude that the normality condition of this test is met.

Secondly, we assume that the variances of both groups are equal. We will define 'equal' to mean that the larger sample standard deviation is not more than twice as large as the smaller sample standard deviation. For this analysis, that ratio is found to be 1.0381968, which is clearly less than 2, so we will assume that the variances are equal across both groups.

Finally, this test assumes independence between observations. In this case, it feels reasonable to say that the number of AID deaths in one state is sufficiently independent of AID deaths in another state. There may be a small level of spatial correlation between neighboring states, however we are assuming that correlation to be neglible because it feels unlikely that the driving habits of people in one state significantly affect the driving habits of people in another state, even if the states share a border.

Because we have confirmed that each of the assumptions have been sufficiently met, we will continue with the analysis. The hypotheses for this test are displayed below.

$$H_o: \ \mu_1 = \mu_2$$

$$H_a: \ \mu_1 > \mu_2$$

Where μ_1 denotes the mean of the average number of AID deaths per state from 2003-2012 and μ_2 denotes the mean number of AID deaths per state in 2018. We will use a significance level of $\alpha = 0.05$.

The results of this one-sided two-sample t test are shown below. With a p-value of 0.22, we fail to reject the null hypothesis. There is not significant evidence that the mean of the 9-year average AID deaths per state is greater than the mean of the number of AID deaths for 2018. The 95% confidence interval, shown in the table below, provides a plausible range of values for the difference in the means of each group. Because our hypothesis test was one-sided, this confidence interval should not be used to reject or fail to reject the null hypothesis of the test we performed. The confidence interval does, however, allow us to make a conclusion about a test with Ha: $\mu_1 neq \mu_2$. Because zero is contained in the interval, we can conclude that there is not evidence that the mean number of AID deaths changed over time from 2003-2012 to 2018.

95% CI	Test-Stat	P-value
(-64.01,145.69)	0.77	0.22

Conclusion

It appears that these data do not provide significant evidence to support the CDC's claim that alcohol impaired driving deaths have signficantly decreased in recent years. Our first comparison group had 50 observations, one for each state, with each observation being the average number of AID deaths in that state for the period 2003-2012. Our second comparison group was simply the number of AID deaths per state in 2018. Because the data had already been aggregated over the 9-year period, it made interpretation of this analysis a bit more complex than a simple comparison of two years, say 2015 and 2018. Even though we did not find that the mean for 2018 was smaller than the mean for 2003-2012, we can still observe that there does seem to be a decreasing trend in the number of AID deaths, which is still a trend in the right direction. Based on these results, we could potentially argue for the need to look into whether current alcohol-impaired driving laws are really taking care of the issue, or at least we could argue for the need to look into the issue

further to see whether AID deaths are truly decreasing. Perhaps they have only been decreasing over the last couple of years, which we did not look at in our analysis. Either way, it is certainly an interesting result.

Opioid Prescription Rates:

One-Way Analysis of Variance

Carly Lundgreen

Introduction and Data Description

Many have heard of the current 'Opioid Crisis' that the United States is currently facing. The U.S. Department of Health and Human Services reports that more than 130 people die from opioid-related drug overdoses every day. But it's not just the general public that face these issues; the National Institute on Drug Abuse reports that more than 1 in 10 veterans who seek care at the U.S. Veteran's Administration meet the criteria to be diagnosed with a substance use disorder (including opioid addiction). This fact becomes the basis for our analysis. We are interested in discovering whether opioid prescription rates differ between different regions of the United States. The U.S. Department of Veteran's Affairs released data opioid prescription rates at 146 Veteran's Affairs (VA) facilities from 2018. We have classified each of these facilities, based on the state they are located in, into one of four U.S. Census-designated regions: Northeast, Midwest, South, and West. Because we plan to analyze differences in the average opioid prescription rate for each of thes regions, we will perform a one-factor analysis of variance to answer our question.

Exploratory Data Analysis

The plots below show a map displaying average opioid prescription rates at VA facilities in 2018 and paired boxplots that compare the distributions of opioid prescription rates by U.S region. From the map, we can see that the average prescription rate in Idaho is over 12.5%, while the average Prescription rates in Texas or Florida, for example, are much lower than that. Figure 2, also shown below, displays. The boxplots show that the West region appears, in general, to have the highest prescription rates in comparison to other regions, while the Northeast appears to have the lowest. When comparing the mean rates for each region, this is certainly the case. The spread of the prescription rates also appears to differ between regions, so we will need to consider this in the next section.

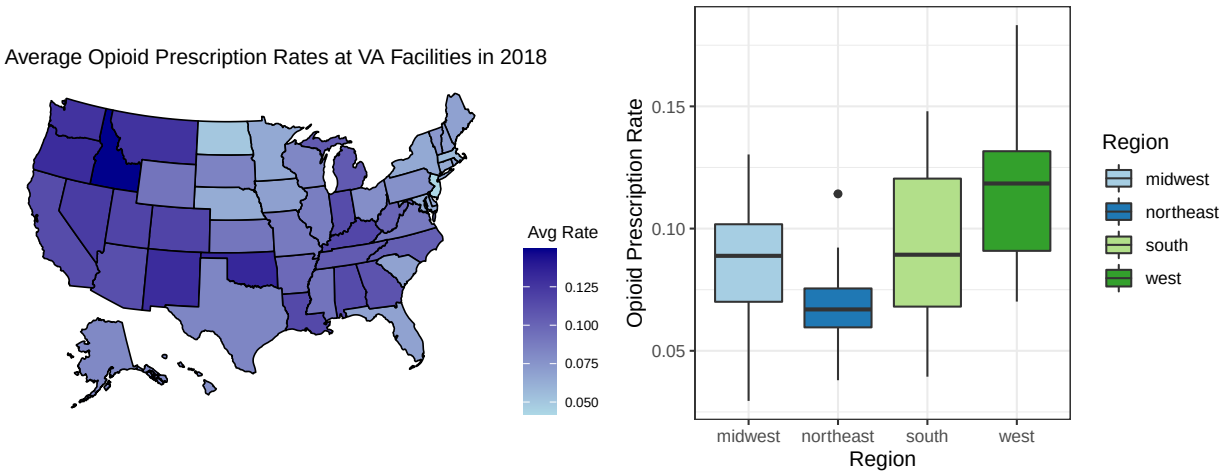


Table 1, displayed below, displays summary statistics of opioid prescription rates for each of the 4 regions. Interestingly, The median rates for the Midwest and South regions are the same (when rounded to three decimal points). We can see that the mean rate is a bit higher for the South, however, which is consistent with the slight right-skewness seen in the boxplot for the South from Figure 2.

Table 1: Summary Statistics of Prescription Rates by Region

region	Mean	SD	Median	IQR	Min	Max
midwest	0.085	0.023	0.089	0.032	0.030	0.130
northeast	0.068	0.017	0.067	0.016	0.038	0.114
south	0.094	0.030	0.089	0.052	0.039	0.148
west	0.116	0.029	0.118	0.041	0.070	0.183

Analysis

We will utilize a one-factor analysis of variance (ANOVA) to compare the mean presecription rates between regions. The hypotheses for this test are as follows:

$$Ho : \mu_{MW} = \mu_{NE} = \mu_S = \mu_W$$

$$Ha : At\ least\ one\ mean\ is\ different\ from\ the\ rest$$

Where each μ is the mean prescription rate at VA facilities in 2018 for the respective region, which is designated by the subscript on μ.

Before proceeding with the analysis, we first need to check the assumptions of a one-way ANOVA. The first assumption that we will check is that the distribution of rates for each region is approximately normal. To test this, we opted to run the ANOVA and simply extract the residuals from the test. If these residuals are sufficiently normal, then the normality condition is met. After extracting the residuals, we performed a Shapiro-Wilk test for normality on the residuals. The null hypothesis of this test is that the sample came from a normally distributed population. The p-value for the Shapiro-Wilk test was 0.2585952, so we fail to reject the null hypothesis and conclude that the condition of normality is met.

The second assumption we checked was that the variances of rates for each region are equal. To check this condition, we utilized Levene's test for homogeneity of variances. The null hypothesis of this test is that population variances are equal (ie the variance of rates are the same for each of the four regions). After conducting the test, we calculated a p-value of 0.003, which is significant at the 0.05 level, indicating that there is likely a difference in variances between groups. This is a potential place for further analysis, where a transformation of our response variable or different definitions of U.S. regions may be considered. For our purposes now, we will proceed with the analysis, keeping in mind this assumption.

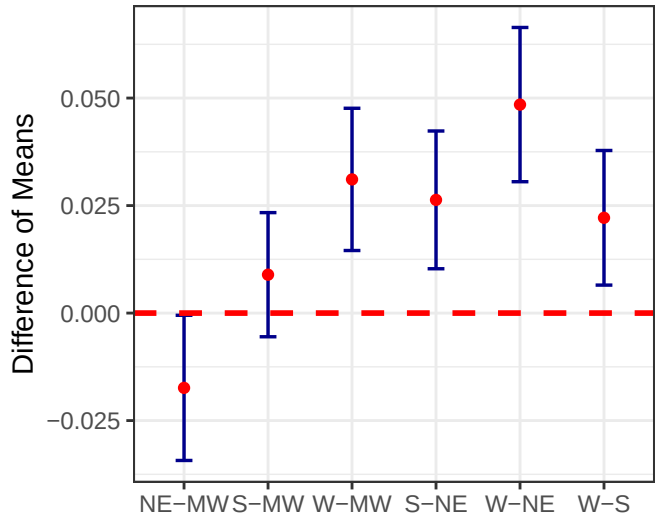
Finally, we consider independence. We believe it is reasonable to assume that prescription rates are independent between VA facilities, so we will assume that independence is a valid assumption here.

The results of our ANOVA test are shown in the table below. We can see that the p-value is less than the significance level of 0.05, so we reject the null and conclude that at least one U.S. region's average opioid prescription rate at VA facilities in 2018 differs from the other regions.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	3	0.03	0.01	17.45	0.0000
Residuals	140	0.09	0.00		

The plot below illustrates 95% confidence intervals for the differences in mean prescription rates for each of the six possible pairs of regions. We can see that the intervals for the difference of mean prescription rates for West-Midwest, West-Northeast, and West-South are all greater than zero, indicating that the mean prescription rate for the West is significantly higher than the other regions. We can make a similar observation about the Northeast-Midwest pairing, where the interval is less than zero. The Midwest region has a significantly higher mean prescription rate than the Northeast. There is not a significant difference in average perscription rate between the South and the Midwest, because the confidence interval contains zero.

This is consistent with our observation made in Table 1, where the median prescription rates for these two regions were the same.



Conclusion

It appears that VA facilities in the West had a higher average prescription rate than VA facilities in the Northeast, South, and Midwest. We also wanted to note that we opted to utilize government-designated regions because the data originated from a government organization. There are likely other regional definitions that would be perfectly reasonable and acceptable, and where we would not run into the problem of unequal variances within groups. The variance assumption leaves this analysis open for further tweaking and exploration. Perhaps a transformation of the rate may help allow us to conclude that the variances are equal across groups. Either way, this analysis does show that opioids are being prescribed at VA facilities all across the United States. With the current opioid addiction crisis in the country, and among veterans, these data take on new meaning, and the relatively high prescription rates in the West open the door to further questions regarding substance abuse control across the country.

Benzodiazepines and Memory:

Two-Way Analysis of Variance

Carly Lundgreen

Introduction and Data Description

Harvard Medical School reported on a study a study performed by a team of researchers from France and Canada which linked benzodiazepine use to an increased risk of being diagnosed with Alzheimer's disease. In the study, the greater people's cumulative dose of benzodiazepines, the higher their risk. However, this was not a randomized clinical trial, so causation could not be established. We are interested in determining whether the effects of these anti-anxiety medications on memory depend on the dosage and type of drug. In other words, we are interested in whether the effect of the type of benzodiazepine on memory depend on the dosage of the medication.

While we do not have data about Alzheimer's disease risk and benzodiazepine usage, we did retrieve a dataset that compares memory test scores for 132 subjects who took a randomly assigned drug every day for one week.This dataset was retrieved from Kaggle, which offers a public data platform off of which we accessed this data. Citations are included in the appendix of this portfolio. The subjects took the memory test before the study period began, and then took the test after the period ended. The scores were recorded for each subject. The two drugs in this study were Alprazolam (Xanax) and Triazolam (Halcion). A placebo was also utilized, but we opted to compare only the two drugs. Each drug had three dosage levels: 1, 2, and 3. For Alprazolam, the levels were, in order: 1mg, 3mg, and 5mg. For Triazolam, the levels 1, 2, and 3 were 0.25mg, 0.5mg, and 0.75mg, respectively. Our response variable for this analysis will be the difference in test score for each subject (after taking the medication minus before taking the medication). A negative difference indicates that the test score decreased after taking the medication, and vice versa.

Exploratory Data Analysis

Figure 1, displayed below, illustrates the distribution of test score differences for combinations of drug and dosage. Note that the drug label "A" indicates Alprazolam and "T" indicates Triazolam. We can see that the distributions for each of these drug/dosage combinations are comparable in spread. The median test score difference for Triazolam appeared to be approximately zero for every dosage level, while the median test scores for Alprazolam appear to differ greatly depending on dosage.

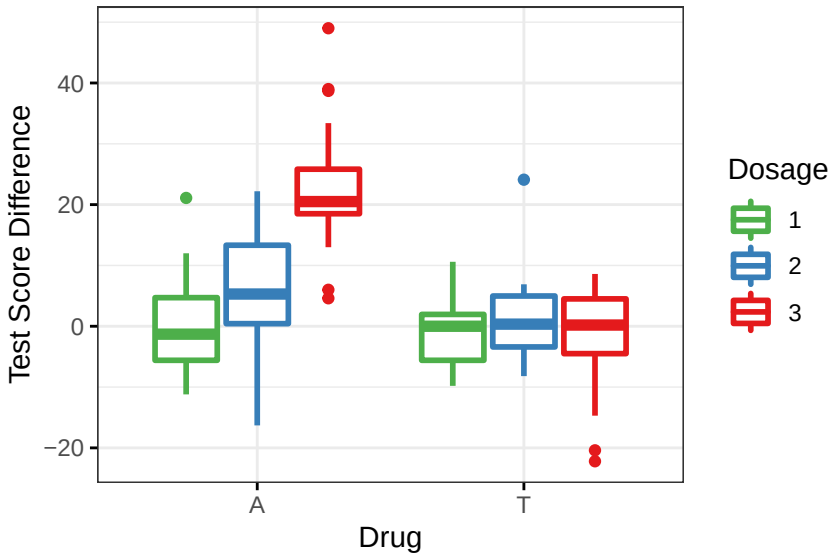


Figure 1: Boxplots comparing test score differences by drug and dosage

Figure 2, below, displays the mean test score difference for each combination of drug and dosage level. It does appear that the anti-anxiety drugs' effects on memory depend on the dose of the respective medication. For example, on dosage level 1 of Alprazolam (Xanax), the after-before difference in test score was just about zero, so there wasn't really an effect on memory in that case. However, the test score difference was very positive for Xanax at dose 3, indicating that the average test score greatly improved after subjects took that drug. Interestingly, Triazolam (Halcion) at dose levels 1 and 3 were associated with a decrease in the average memory test score after taking that drug at those dose levels, while dose level 2 appears to be associated with an increase in test score. Perhaps Triazolam has an 'ideal' dose range, where doses outside of that range are associated with decreased memory recall. This is in contrast to Alprazolam, which seems to have positive effects no matter the dose range, but an extremely positive effect at the highest dose level.

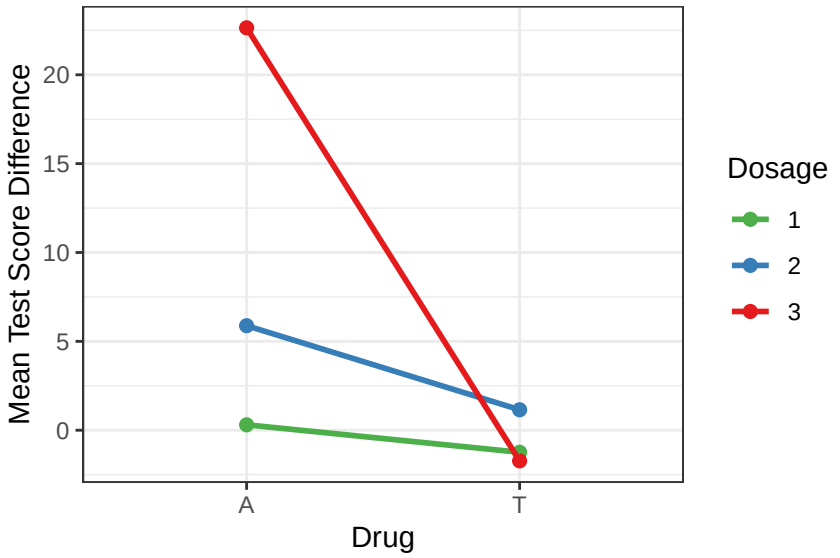


Figure 2: Plot of the mean test score differences for each drug/dosage combination

Table 1 below illustrates summary statistics for the difference in test scores by drug and dosage level. The

group that obviously stands out the most within this table is the A:3 group, which is Alprazolam at Dosage level 3 (5mg). The mean test score difference is 22.641, which is far higher than the rest of the groups. We observed this in the interaction plot above. The difference between this group and the rest is so high that as a point of further analysis we might consider returning to the data source to make sure there was not a misstype in the dataset. For now, we assume that the data measurement is correct and proceed with the analysis.

Drug	Dosage	Count	Mean	SD	Median
A	1	23	0.304	8.277	-1.30
A	2	22	5.882	9.210	5.30
A	3	22	22.641	10.405	20.50
T	1	22	-1.241	5.530	0.00
T	2	22	1.150	6.718	0.35
T	3	21	-1.724	8.768	0.20

Analysis

We want to find out if there is a statistically significant interaction between `Drug` and `Dosage` on the difference in memory recall test score. To do this, we will utilize a two-factor analysis of variance (ANOVA). We will test the following hypotheses:

$$H_o: \textit{ There is no interaction between drug and dosage level}$$
$$H_a: \textit{ There is an interaction between drug and dosage level}$$

Before proceeding with the test, however, we will check that the assumptions of the two-way ANOVA test are met. First, we will make sure that the normality condition is met. This condition requires that the distributions of test score differences are normally distributed for each group. Because we have six different drug/dosage combinations, we will instead fit our ANOVA model and look at the distribution of residuals from the ANOVA. A histogram of the residuals is displayed below in Figure 3. This histogram certainly appears normal, but we will utilize a Shapiro-Wilk test for normality to confirm our assumption. The p-value from the Shapiro-Wilk test was found to be 0.0920689, which is not significant at the 0.05 level, so we fail to reject the null hypothesis that the residuals are normally distributed. The normality assumption appears to be sufficiently met.

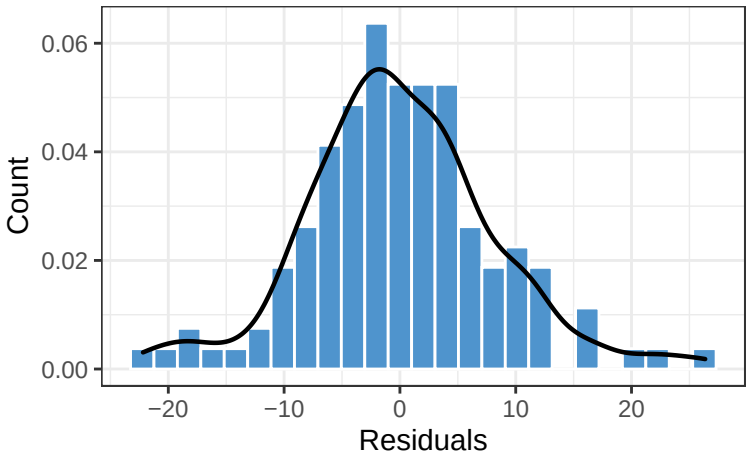
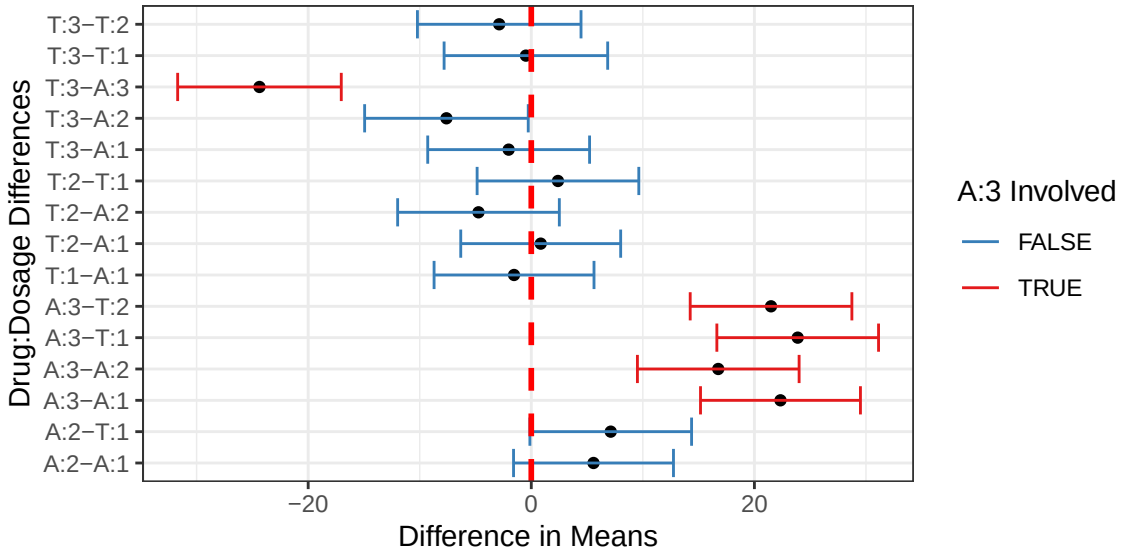


Figure 3: Histogram of Residuals from ANOVA test

Secondly, we check the condition of equal variaince across comparison groups. We utilized Levene's test for homogeneity of variance to determine whether the variance of test score differences is the same across drug/dosage groups. This test returned a p-value of 0.476, which led us to fail to reject the null hypothesis that the variances are homogenous across groups. We believe the assumption of equal variance is sufficiently met, so we finally check the assumption of independence, which requires that the memory test score differences are independent from subject to subject. Because the test was given in controlled conditions, we believe that each person's test score was independent of the rest. Because we have confirmed that the assumptions have been met for this scenario, we will finally report the results of the ANOVA test in the table below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Drug	1	3337.52	3337.52	48.39	0.0000
Dosage	2	2808.76	1404.38	20.36	0.0000
Drug:Dosage	2	3326.03	1663.02	24.11	0.0000
Residuals	126	8689.56	68.96		

We can see that the p-value for the interaction between Drug and Dosage is effectively zero, which is obviously significant at the 0.05 level, so we conclude that there is a significant interaction between Drug and Dosage on the difference in memory test scores. The effects are also significant for Drug and Dosage alone, with p-values effectively zero for each of the main effects. To analyze which Drug/Dosage combinations are contributing to this interaction effect, we will look at pairwise confidence intervals, shown in the plot below.



We can see that every interval where the A:3 interaction was involved, there was a significant difference in mean test score. All the other confidence intervals contain zero, indicating that the difference in mean test score was not significant. These are the intervals displayed in red. The Alprazolam 5mg combination appears to be the source of the significant interaction between drug and dosage.

Conclusion

As stated above, it appears that the Alprazolam at 5mg was the major factor level combination leading us to conclude that there is a significant interaction between drug and dosage on memory recall test scores. As mentioned previously, we would be interested to find out whether there was a misstype in the data, or if the difference in test scores were really that large for those subjects taking 5mg of Alprazolam. If the data was correct, then Xanax leads to an incredibly large boost of memory after taking 5 mg for a week. At other

drug/dosage level combinations, it doesn't appear that there were was much of a difference in test score after taking the drug for a week. We would probably recommend another study design that has subjects take the drug for a longer period, as most anti-anxiety/anti-depressants have side effects that take 1-3 weeks (or more) to "settle down" after beginning to take the medication for the first time. These side effects could have contributed to the memory test scores. A final note about the data–we didn't have much information about the subjects that participated in the study. The only information given was that they were 'islanders'. This could have many meanings, and it also prevents us from generalizing any of these results beyond the islander population. We also recognize that only two benzodiazepines were considered in this study, so it is very likely that other benzodiazepines have differing, and likely more significant, effects on memory.

Happiness and Alcohol Consumption:

One-Factor Analysis of Covariance

Carly Lundgreen

Introduction and Data Description

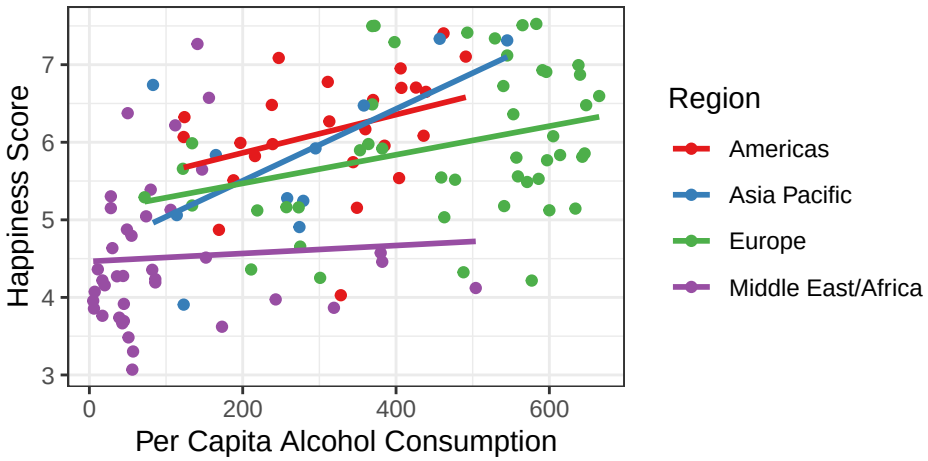
In 2016, the World Happiness Report rated Denmark as the happiest country in the world by considering variables such as log GDP per capita, healthy life expectancy at birth, freedom to make life choices, and perceptions of corruption. Because happiness is so subjective, and dependent upon likely innumerable different variables, it is notoriously difficult to measure. However, these studies of global happiness can reveal many truths about the state of the country in question and the quality of life of the people living there, regardless of how close scientists get to the "true model" of happiness.

We are interested to find out whether where you live in the world affects how happy you are. We can already infer, based on the 2016 World Happiness Report and subsequent reports (where Denmark consistently ranks among the happiest in the world), that where you live does affect your happiness level. In addition to location, however, we want to consider the effect of alcohol consumption on happiness. In other words, we want to see how world region affect happiness scores if we control for alcohol consumption. The dataset utilized in this analysis was retrieved from Kaggle, which offers a public data platform from which we accessed this data. Citations are included in the appendix of this portfolio.

Our dataset reports the average happiness score for 122 countries in the world from the year 2016. These countries were already segmented into 9 different regions (North America, Latin America/Caribbean, Eastern Asia, Southeastern Asia, Australia/New Zealand, Central/Eastern Europe, Western Europe, Middle East/North Africa, and Sub-Saharan Africa). In order to simplify the analysis, we aggregated these nine regions into just four: Americas, Asia Pacific, Europe, and Middle East/Africa. Our response variable, happiness score (`score`), was the average response of a survey of people from each country to the question "How happy are you on a scale of 1-10?" The factor, again, was `region` with 4 levels. Finally, we considered alcohol consumption as our covariate that we will control for in this analysis. There were three variables, Beer per capita, Spirit per capita, and Wine per capita, that we summed into a single variable, alcohol per capita (`alcohol_pc`). We were not able to find out whether these were the numer of beverages per capita consumed per year/month/week, so we will just consider this variable as the number of alcoholic beverages consumed per year (where 'alcoholic beverages' refers to beer, spirit, or wine). Our goal is to identify differences in average happiness score between regions, while controlling for alcohol consumption. To accomplish this, we will perform a one-factor analysis of covariance (ANCOVA) test.

Exploratory Data Analysis

The plot below displays a scatterplot of alcohol per capita by happiness score with regression lines drawn for each of the four regions that we defined during the data cleaning process. It appears that the slopes of each of these lines are positive, indicating that alcohol consumption, on average, is associated with higher happiness scores for each of the regions. The slopes are different for each region, however, indicating a differing effect for each region. For example, it appears that alcohol does not have very much of an effect on happiness in the Middle East/Africa.



Analysis

We will utilize a one-way analysis of covariance (ANCOVA) to identify whether the mean happiness score differs between region after considering the effect of alcohol consumption. The hypotheses we will utilize are as follows:

Hypotheses:

$$Ho : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$Ha : At\ least\ one\ mean\ is\ different\ from\ the\ rest$$

Where μ_i (i = 1,2,3,4) are the mean happiness scores after adjustment for per-capita alcohol consumption for the following regions, respectively: Americas, Asia Pacific, Europe, and Middle East/Africa.

Before proceeding, we need to consider the assumptions of the one-factor ANCOVA procedure. First, we will identify whether there is a significant interaction between region and alcohol consumption. The results of an F test to determine whether there is a significant interaction between these two variables is displayed in the table below.

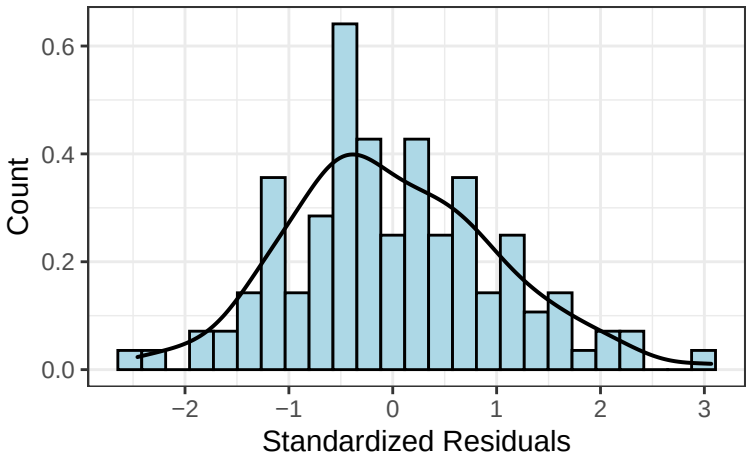
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	3	59.40	19.80	25.20	0.0000
alcohol_pc	1	8.05	8.05	10.24	0.0018
region:alcohol_pc	3	2.65	0.88	1.12	0.3428
Residuals	114	89.57	0.79		

We can see from the above table that there is not a significant interaction between yearly alcohol consumption per capita and region of the world. This means that we will not have to consider an interaction between these variables in our ANCOVA procedure.

The next assumption we will make is that relationships are linear. We can see from the scatterplot above that, for each region, it is reasonable to say that alcohol consumption and happiness have a linear relationship. Next, we assume that happiness scores are destributed normally for each region. Instead of looking at all four regions, we will instead fit a linear model, including `alcohol_pc` as a covariate, and determine whether the residuals are normally distributed. The model is shown below:

$$score = \alpha + \beta_{alcohol} + \beta_{region} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

A histogram of the residuals from this model is displayed below.



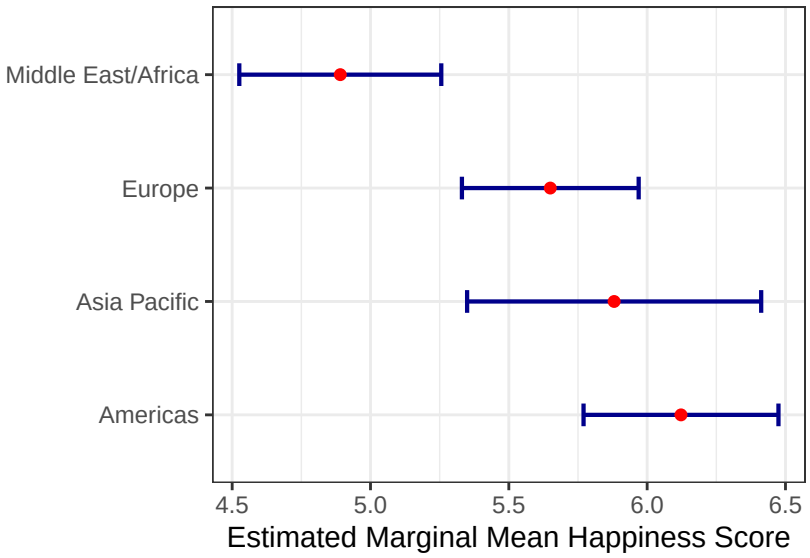
In addition to the above plot, which appears normally distributed, we conducted a Kolmogorov-Smirnov test for normality, which resulted in a p-value of 0.7163945, leading us to fail to reject the null hypothesis that the residuals are distributed normally. The final assumption of our ANCOVA procedure is a constant variance of happiness scores between regions. Equivalently, we can look at the variances of the residuals from the model that includes alcohol consumption as a covariate (i.e. controls for the effects of alcohol consumption) with respect to region. We utilized Levene's test for homogeneity of variance, which returned a p-value of 0.453, leading us to fial to reject the null hypothesis that the variances of scores for each region are equal. We can confirm that the conditions for the ANCOVA procedure are met.

We fit an ANOVA model with type III sums of squares (because we want to control for the effect of alcohol consumption) in order to evaluate the difference in happiness score between regions. The results are shown below.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	364.26	1	462.15	0.0000
alcohol_pc	8.05	1	10.21	0.0018
region	19.60	3	8.29	0.0000
Residuals	92.22	117		

The covariate, alcohol per capita, was significantly related to the happiness score after controlling for the effect of region. However, we are interested in the effect of region on happiness score. There was also a significant effect of region on the happiness score after controlling for the effect of alcohol consumption. In other words, after adjustment for alcohol consumption per capita, there was a statistically significant difference in happiness score between the groups. In order to analyze where these differences are taking place in the model, we will consider the marginal means for each region. The table below shows the marginal mean happiness score, along with associated 95% interval lower and upper bounds, for each region. We can see that the Americas had the highest marginal mean happiness score, while the Middle East/Africa had the lowest. These means, and their associated intervals, are plotted below.

Region	Est. Marginal Mean	Lower Bound	Upper Bound
Americas	6.122	5.770	6.475
Asia Pacific	5.881	5.349	6.412
Europe	5.650	5.330	5.969
Middle East/Africa	4.891	4.526	5.256



We can see that the mean happiness score for the Middle East/Africa is significantly lower than all the other scores, as its 95% confidence interval does not overlap the intervals for any of the other regions. Furthermore, we wanted to analyze the pairwise differences of these marginal means and associated significance. The table below shows the difference in marginal means for each possible regional pair. Again we can see that the significant differences of marginal means are only those pairs that involve the middle East/Africa.

Contrast Pair	Difference	P-Value
Americas - Asia Pacific	0.242	0.877
Americas - Europe	0.473	0.192
Americas - (Middle East/Africa)	1.232	0.000
Asia Pacific - Europe	0.231	0.887
Asia Pacific - (Middle East/Africa)	0.990	0.013
Europe - (Middle East/Africa)	0.759	0.046

Conclusion

Based on our analysis of covariance, we conclude that the mean happiness score for the Middle East/Africa is significantly lower than the mean happiness scores for the Americas, Europe, and the Asia Pacific region, when controlling for the effect of alcohol consumption on happiness. In other words, where you live in the world has a significant effect on happiness score when controlling for the effects of alcohol consumption. However, we also acknowledge that there are many, many other possible variables that contribute to happiness levels that could change or alter these results if they were included in our analysis. Regardless, our results support those of the World Happiness Report: average happiness levels differ across the world. Though we also note that happiness is of itself a personal endeavor, and happiness can usually be found regardless of circumstance.

Appendix

Applications and the Linear Model

Each of these 5 applications is a special case of the linear model $y = x\beta + \epsilon$

Analysis of One Mean

This example of a one-sample t procedure is a special case of the linear model if we specify y as a vector of the difference in yellow-fever antibody concentrations, and J as a vector of 1s, the same dimension as the y-vector (nx1). Beta-hat, the vector of coefficient estimates, simplifies to equal the sample mean concentration of yellow fever antibodies (ybar). and the sample variance simplifies to the sum of the squared deviations of each concentration from the sample mean concentration divided by n - 1, where n = 587 in this scenario. With these specifications, the null hypothesis $\mu = 0.5$ can written as beta-hat = 0.5. The one-sample t test procedure follows, simplifying to $t = (ybar - 0.5)/sqrt(s^2/n)$.

Analysis of Two Means

This application of a two-sample t procedure is a special case of the linear model if we specify y as the 100x1 vector of deaths to due alcohol impaired driving for each state, with the first 50 values being the average deaths for each of the 50 states in 2003-2012, and the next 50 values being the average deaths for each of the 50 states in 2018 (all due to AID). Then, we define the X matrix as a 100x2 matrix with 50 ones in the first column, followed by 50 zeroes in the first column. Then, in the second column, 50 zeroes followed by 50 ones. This is to designate which group each observation belongs to. Using these specifications, beta-hat simplifies to a 2x1 vector with the sample mean of the average deaths due to AID in 2003-2012 in the first row, and the sample mean deaths due to AID in 2018 in the second row. The sample variance simplifies to the typical 'pooled sample variance' equation seen in introductory statistics courses. The hypothesis test $\mu_1 = \mu_2$ can be written as Ho: 1 -1 = 0. The t-test formula will be written as \$t = c*betahat / sqrt((s^2)C((X'X)^-1)C') \$.

One-Factor ANOVA

This application of a one-way ANOVA can be written as a special case of the linear model by writing the y vetor as the opioid precsription rates for each of the VA facilities in 2018. Then the X matrix is defined based on the basis function expansion where the first column is 1 if midwest, 0 if not. The second column is 1 if northeast, and 0 if not. The third column is 1 if south, 0 if not. And the 4th column is 1 if west, 0 if not. Then betahat simplifies to the following: beta_0 = mean opioid prescription rate for northeast region; beta_1 = mean rate for midwest - mean rate for south; beta_2 = mean rate for northeast - mean rate for south; beta_3 = mean rate for west - mean rate for south. Then we can define C = [1 0.25 0.25 0.25] and utilize the formula for F with contrasts.

Two-Factor ANOVA

This application of a two-factor ANOVA can be written as a special case of the linear model by modifying the one-factor ANOVA specifications. Instead of just 1 factor in the basis function expansion, we now have to consider two factors. This will cause the X matrix to be much larger. In this scenario, our first factor had two levels: A and T. Our second factor had 3 levels: dosages 1,2,and 3. With this ANOVA, we are testing whether there is an interaction between drug and dosage. We will have to have columns of 1s and 0s to capture every combination of drug and dosage, and then we can use the F forula that employs C*betahat to define the test statistic.

One-Way ANCOVA

This application of a one-factor ANCOVA is essentially blending ANOVA and linear regression. We are including the covariate, alcohol consumption, as a quantitative variable in our analysis, and the factor variable region as the group variable. Essentially, we are fitting the following model: $Y_{ij} = \mu + \tau_i + B(x_{ij} - xbar) + \epsilon_{ij}$, where Y_{ij} is the jth observation in the ith categorical group (region). τ_i is defined as the effect of the ith level of region, μ is the grand mean, and xbar is the global mean for the covariate (alcohol per capita). Finally, ϵ_{ij} is the associated unobserved error term for the jth observation in the ith region. The $tau_i's$ will all sum to zero, and we used the standard assumptions of the linear model.

Data Citations (URLs)

Domingo, Cristina, & Mejía, José E. (2020). Long-term immunity against yellow fever in children vaccinated during infancy: a longitudinal cohort study (Version 1.0.1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3333025

https://www.data.va.gov/d/dwpj-hgp7?category=dataset&view_name=Opioid-Prescribing-Rates-at-VA-Facilities-2012-201

https://www.kaggle.com/steveahn/memory-test-on-drugged-islanders-data

https://www.kaggle.com/marcospessotto/happiness-and-alcohol-consumption