PH717 Team 2 Project

Pallavi Goel, Adam Deutsch, Dana Vasers, Matthew Zuniga, Mahlet Kelecha

Project Component 1

Dataset: Water and sanitation in Zambia

Research question: Do households involved in the ZSHP Project have improved sanitation?

Primary exposure of interest: ZSHP initiation (dichotomous)

Primary outcome of interest: Improved sanitation (dichotomous)

Hypothesized direction of the association: If households are involved in the ZSHP Project, they are expected to have improved sanitation.

Project Component 2

Methods: The study design consists of two cross-sectional surveys comparing participation in the ZSHP program and improved sanitation. The length of follow-up was 3 years, and the number of participants was 1,170. 723 participants were initiated in the program and 447 participants were not initiated in the program. Participant surveys were conducted randomly among those living in 1 of 8 provinces in rural Zambia between 2012 and 2017.

Results:

	Improved Sanitation	**No Improved Sanitation**	**Total**
Initiation	598	125	723
No Initiation	340	107	447
Total	938	232	1170

Risk ratio = CI_E/CI_U = (598/723)/(340/447) = 1.09

Those initiated in the ZSHP program had 1.09 times the risk of improved sanitation than those who were not initiated in the program. This is based on data collected from 8 provinces in rural Zambia between 2012 and 2017.

R-Code:

```
attach(WASH_data)
table(ZSHP_initiation, improved_sanitation)
```

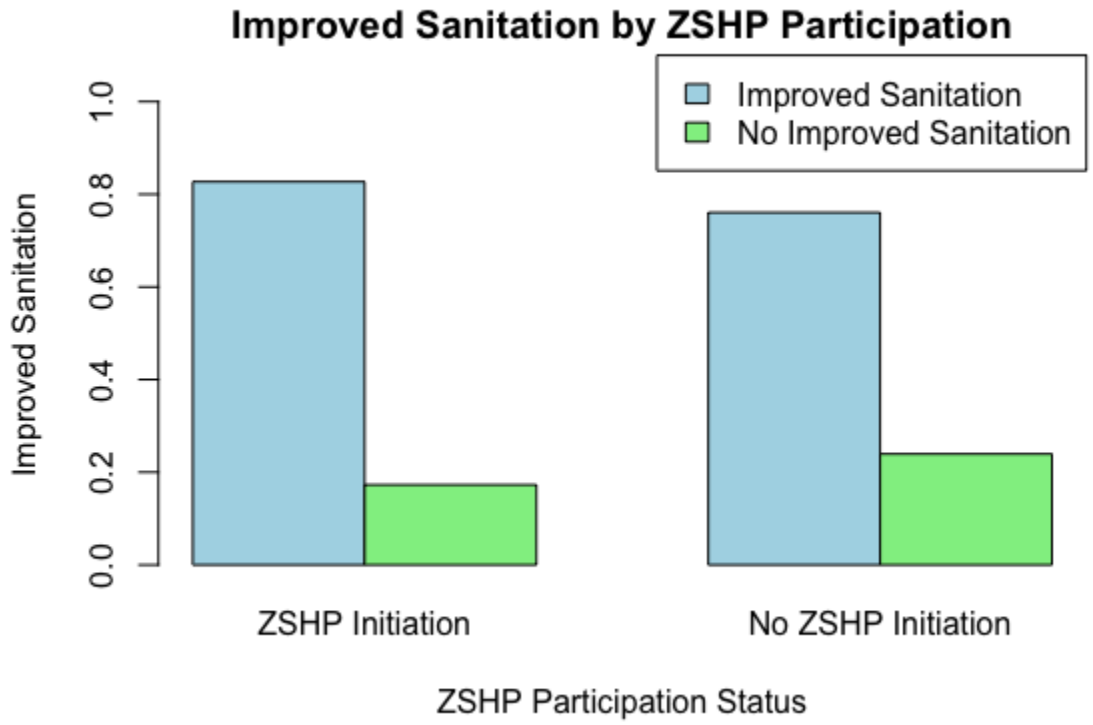
Project Component 3

Table 1. Baseline characteristics by exposure category

	Exposure Category: ZSHP Initiation	**Exposure category: No** ZSHP Initiation
Sample size: n	**723**	**447**
Average age of the head of the household: Mean (sd)	**38.76 (10.15)**	**37.85 (9.66)**
Average number of persons living in the household: Mean (sd)	**6.17 (2.51)**	**5.87 (2.49)**
Highest education of the head of the household: frequency (%) **1 = No education** **2 = Primary** **3 = Secondary** **4 = Tertiary**	**59 (8.16%)** **386 (53.39%)** **270 (37.34%)** **8 (1.11%)**	**46 (10.29%)** **247 (55.26%)** **145 (32.44%)** **9 (2.01%)**
Wealth quantile: frequency (%) **1 = Poorest 20%tile** **2 = Second 20%tile** **3 = Third 20%tile** **4 = Fourth 20%tile** **5 = Wealthiest 20%tile**	**115 (15.9%)** **130 (17.98%)** **142 (19.64%)** **164 (22.68%)** **172 (23.79%)**	**120 (26.85%)** **70 (15.66%)** **128 (28.64%)** **60 (13.42%)** **69 (15.44%)**

Results:
Many characteristics between the two exposure groups are similar such as average age of the head of household (38.76 yrs vs 37.85 yrs), average number of persons living in the household (6.17 ppl vs 5.87 ppl) and the distribution of highest education achieved. The highest percentage of education in both samples was primary education, with 53.39% of the initiated sample group and 55.26% of the not initiated sample group. The sample size of those who were initiated into ZSHP (723) is larger than the sample size of those not initiated (447). In addition, there are higher frequencies of higher wealth quantiles among those who were initiated in the program (Wealthiest 20% -23.79%) compared to those who were not (Wealthiest 20%- 15.44%).

Figure 1.



Results: In those households that were initiated into the ZSHP Program, there was a higher frequency of improved sanitation versus no improved sanitation. In addition, the overall frequency of improved sanitation in those initiated was 82%, compared to 76% in the group that was not initiated. There is a higher frequency of no improved sanitation in the not initiated group (24%) compared to the initiated group (18%), as well.

R-Code Component #3:

```
table(ZSHP_initiation)
tapply(hh_size,ZSHP_initiation,mean)
tapply(hh_size,ZSHP_initiation,sd)
tapply(age_hh,ZSHP_initiation,mean)
tapply(age_hh,ZSHP_initiation,sd)
table(ZSHP_initiation, educ_hh)
table(ZSHP_initiation, wealth)

barplotimprovedsanitation <- prop.table(table(improved_sanitation,ZSHP_initiation), 2)
barplot(barplotimprovedsanitation,
      main= "Improved Sanitation by ZSHP Participation",
      ylab = "Improved Sanitation",
      ylim = c(0, 1.0),
      xlab="ZSHP Participation Status",
      col=c("light blue","light green"),
```

```
      names.arg=c("ZSHP Initiation", "No ZSHP Initiation"),
      legend = c("Improved Sanitation", "No Improved Sanitation"), beside=TRUE, args.legend =
list(x = "topright", inset = c(0, -0.1)))


prop.table(table(improved_sanitation,ZSHP_initiation), 2)
                ZSHP_initiation
improved_sanitation         1         2
                1 0.8271093 0.7606264
                2 0.1728907 0.2393736
```

Project Component 4

Table 2. Outcomes by exposure and comparison analysis

	Exposure: ZHSP Initiation	Exposure: No ZHSP Initiation	Risk difference (p-value chi-square test no difference)	Risk Ratio (95% CI; p-value RR = 1)	Odds Ratio (95% CI; p-value RR = 1)
Sample size: n = 1170	723	447			
Risk of Improved Sanitation (Unadjusted)	0.827	0.761	0.066	1.38 (1.10, 1.74) p-value = 0.006	1.51 (1.13, 2.01) p-value = 0.006
Risk of Improved Sanitation (Adjusted)					

Among households that were initiated into the ZHSP sanitation program, they were 1.38 times the risk of improved sanitation in comparison to households that were not initiated into the ZHSP sanitation program. A chi square test was tabulated to observe the difference in risk of improved sanitation depending on program initiation status ($X^2 = 7.68$, p-value = 0.006). A 95% confidence interval was calculated to determine if the true risk ratio lied between the interval (1.10, 1.74); program initiation was observed to be statistically significantly associated with improved sanitation (p-value = 0.006), therefore, we reject the null hypothesis.

R-Code Component #4:

```
> chisq.test(ZSHP_initiation, improved_sanitation, correct = FALSE)
        Pearson's Chi-squared test
data:  ZSHP_initiation and improved_sanitation
X-squared = 7.68, df = 1, p-value = 0.005584
library(epitools)
> riskratio.wald(ZSHP_initiation, improved_sanitation)
$data
       Outcome
```

```
Predictor   1   2 Total
   1     598 125   723
   2     340 107   447
   Total 938 232  1170
$measure
        risk ratio with 95% C.I.
Predictor estimate    lower    upper
        1 1.000000       NA       NA
        2 1.384537 1.100477 1.741919
$p.value
        two-sided
Predictor  midp.exact fisher.exact  chi.square
        1          NA          NA          NA
        2 0.006055803  0.006544889 0.005583664
$correction
[1] FALSE
attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
> oddsratio.wald(ZSHP_initiation, improved_sanitation)
$data
        Outcome
Predictor   1   2 Total
   1     598 125   723
   2     340 107   447
   Total 938 232  1170
$measure
        odds ratio with 95% C.I.
Predictor estimate    lower    upper
        1 1.000000       NA       NA
        2 1.505553 1.126055 2.012948
$p.value
        two-sided
Predictor  midp.exact fisher.exact  chi.square
        1          NA          NA          NA
        2 0.006055803  0.006544889 0.005583664
$correction
[1] FALSE
attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
```

Project Component 5

Confounders: Wealth and Number of People in Household
sanitation = 0.737 + 0.049(ZSHP_initiation) - .008(low_hs) + .009(high_wealth)

Table 2. Unadjusted and Adjusted Outcomes by exposure and comparison analysis

	Exposure: ZHSP Initiation	Exposure: No ZHSP Initiation	Risk difference (p-value chi-square test no difference)	Risk Ratio (95% CI; p-value RR = 1)	Odds Ratio (95% CI; p-value RR = 1)
Sample size: n = 1170	723	447			
Risk of Improved Sanitation (Unadjusted)	0.83	0.76	0.07	1.38 (1.10, 1.74; p-value = 0.01)	1.51 (1.13, 2.01; p-value = 0.01)
Risk of Improved Sanitation (Adjusted for Wealth & # People in Household)			0.05		1.355 (1.007, 1.821; p-value = 3.072e-05)

Among households that were initiated into the ZHSP sanitation program, they had 1.355 times the adjusted odds of improved sanitation in comparison to households that were not initiated into the ZHSP sanitation program. A 95% confidence interval was calculated to determine if the true adjusted odds ratio lied between the interval (1.007, 1.821); program initiation was observed to be statistically significantly associated with improved sanitation (p-value = 3.072e-05), therefore, we reject the null hypothesis.

Based on the adjusted results, wealth and number of people per household were confounders that were successfully adjusted for.

Discussion/Conclusion:

One potential source of bias is how exposure to the ZSHP initiation is measured as the variable description just requires that the household was in a community that had ZSHP project involvement. This could have led to insufficient data collected by the surveys as it doesn't determine if the specific household was involved in the project. This is a non-differential measurement error and likely underestimated the true measure of association.

One potential source of confounding is wealth as it is an independent risk factor for improved sanitation, it is associated with the ZSHP initiation and it's not an intermediate on the pathway between the exposure to the program and improved sanitation. Wealth is a confounder as it can lead to sanitation improvements through the purchase of sanitation materials or updating equipment.

Two limitations of this study are there is no baseline survey included in the data from before the initiation was put in place and the sample size only being 1,170. The dataset we have access to only included the second survey which was from after the initiation occurred. This is an issue as many of the outcome variables are based on improvement such as the one we are looking at which is improved sanitation. Without the baseline for what the sanitation was like before the initiation it can be hard to determine if there was actual improvement in sanitation post-initiation. The sample size is also an issue as only having 1,170 individuals in the study is quite small considering this an intervention being evaluated for use on a larger scale in this region. Having a larger sample size would create more accurate results for the measures and have smaller margins of errors.

The results are not generalizable on a global scale but could be generalizable to other areas similar to rural Zambia such as other rural regions in sub-Saharan Africa. This is based on the eligibility criteria of the study which was individuals living in rural Zambia which is a unique location as shown by the coding of the variables which vary greatly such as water source. People living in rural sub-Saharan countries often face challenges finding regular access to water which may not be as urgent of an issue in other parts of the world where sanitation issues occur. Based upon our data and results the program initiation did lead to improved sanitation and could be a useful tool to be used in the rural sub-Saharan region. Implementing this system into an intervention and regular practice for rural regions in the sub-Saharan region would be useful in preventing issues like infections from poor WASH conditions, reducing stunting, and improving education quality which are the main goals of the program.

R Code:

```
# Dummy variables from continuous, dichotomous, and categorical variables
> Initiation_new <- ifelse(ZSHP_initiation == 1, 1, 0)
> Improved_new <- ifelse(improved_sanitation == 1, 1, 0)
>  low_HS <- ifelse(hh_size < 6, 1, 0)
> high_wealth <- ifelse(wealth > 3, 1, 0)

#Unadjusted Regression
unadj_reg <- lm(Improved_new~Initiation_new)
> summary(unadj_reg)

Call:
lm(formula = Improved_new ~ Initiation_new)
Residuals:
   Min     1Q  Median     3Q    Max
-0.8271  0.1729  0.1729  0.2394  0.2394

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.76063    0.01881  40.432  < 2e-16 ***
Initiation_new  0.06648    0.02393   2.778  0.00556 **
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3977 on 1168 degrees of freedom
Multiple R-squared: 0.006564,	Adjusted R-squared: 0.005714
F-statistic: 7.718 on 1 and 1168 DF, p-value: 0.005556

#Multiple Linear Regression

```
> reg_sanitation <- lm(Improved_new~Initiation_new+low_HS+high_wealth)
> summary(reg_sanitation)

Call:
lm(formula = Improved_new ~ Initiation_new + low_HS + high_wealth)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8811  0.1189  0.1757  0.2211  0.2703

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.737252   0.023812  30.961  < 2e-16 ***
Initiation_new  0.049212   0.024198   2.034   0.0422 *
low_HS         -0.007588   0.023409  -0.324   0.7459
high_wealth     0.094642   0.024176   3.915 9.58e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3954 on 1166 degrees of freedom
Multiple R-squared:   0.02,    Adjusted R-squared:  0.01748
F-statistic: 7.933 on 3 and 1166 DF,  p-value: 3.072e-05
```

 # Create Model

```
 > LogitModel <-glm(Improved_new~Initiation_new + low_HS + high_wealth,family =
binomial(link = logit))
```

#Odds Ratio

```
> exp(LogitModel$coefficients)
  (Intercept) Initiation_new         low_HS    high_wealth
    2.7740987      1.3553314      0.9501458      1.8974030
```

 # CONFIDENCE INTERVAL

```
> exp(confint(LogitModel))
Waiting for profiling to be done...
                 2.5 %   97.5 %
(Intercept)    2.1016344 3.693025
Initiation_new 1.0068598 1.821806
```

low_HS 0.7089528 1.273957
high_wealth 1.3796390 2.635597