HW 03 - Missing Data Assignment

Kyle Barisone

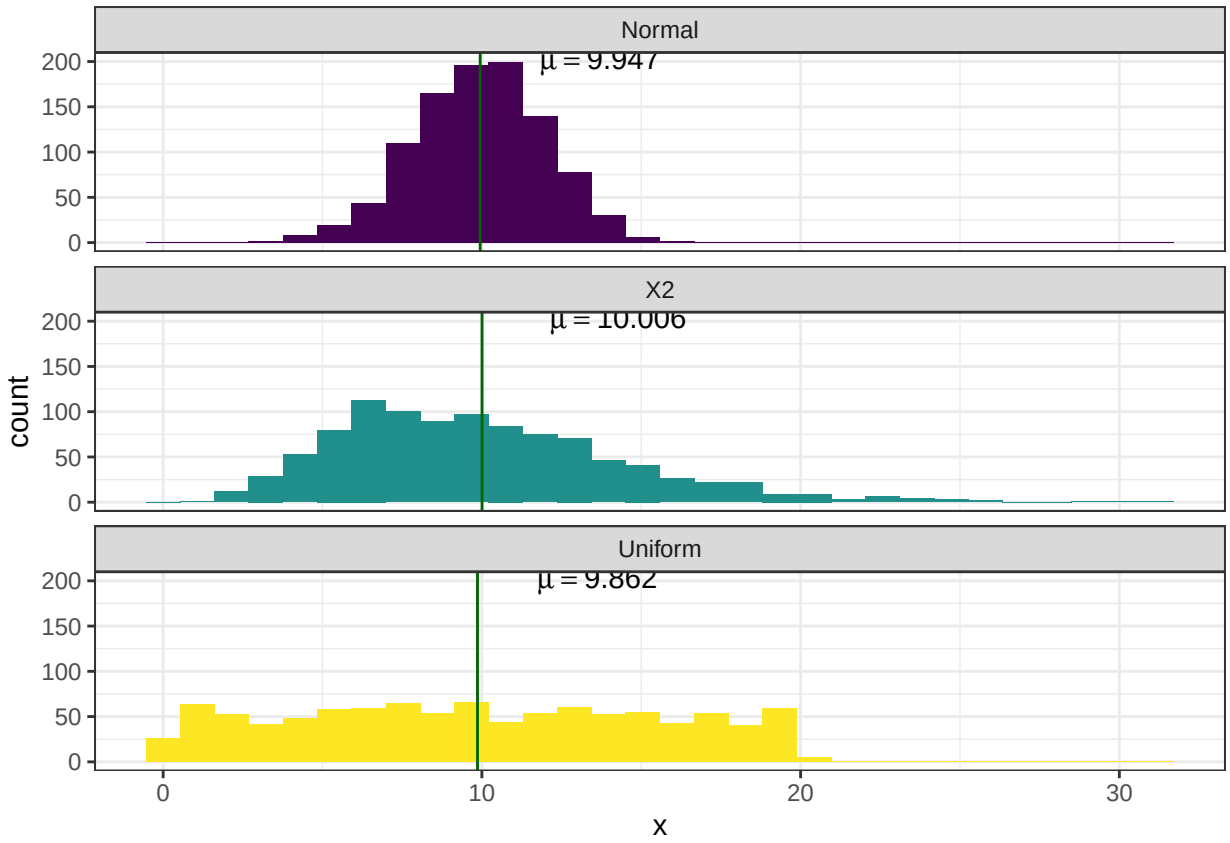
Understanding impact of missing data on parameter estimates.

1. Simulate and describe the following population distributions.

```
set.seed(8675309)
N=1000

# sample data from known/named distributions.
norm.pop.data <- data.frame(x=rnorm(n=N, mean=10, sd=2), dist =
  "Normal")
chi2.pop.data <- data.frame(x=rchisq(N,10), dist = "X2")
unif.pop.data <- data.frame(x=runif(N, 0,20 ), dist = "Uniform")

# combine into one data set
pop.data <- rbind(norm.pop.data, chi2.pop.data, unif.pop.data)

# calculate grouped summary stats
pop.ss <- pop.data %>% group_by(dist) %>%
                    summarise(mean=mean(x), var=var(x),
                              min=min(x), max=max(x))
kable(pop.ss, digits=3) %>% kable_styling(full_width=FALSE)
```

dist	mean	var	min	max
Normal	9.947	4.179	3.101	16.826
X2	10.006	20.293	0.636	31.217
Uniform	9.862	32.403	0.040	19.998

```
# plot pop data - one panel per distribution.
ggplot(pop.data, aes(x=x, fill=dist)) + geom_histogram()  +
      geom_vline(data=pop.ss, aes(xintercept=mean), col="darkgreen") +
      geom_text(data=pop.ss, parse=TRUE, hjust=-.5,
                aes(y = 200, x=mean, label= paste("mu == ", round(mean,3)))) +
      facet_wrap(~dist, ncol=1) + scale_fill_discrete(guide=FALSE)
```



The normal distribution has a mean of 9.94 and a variance of 4.18. The data ranges from 3.1 to 16.83 and resembles the shape of a normal distribution.

The chi square distribution has a mean of 10.01 and a variance of 20.29. The data ranges from .64 to 31.22 and is properly skewed to the right.

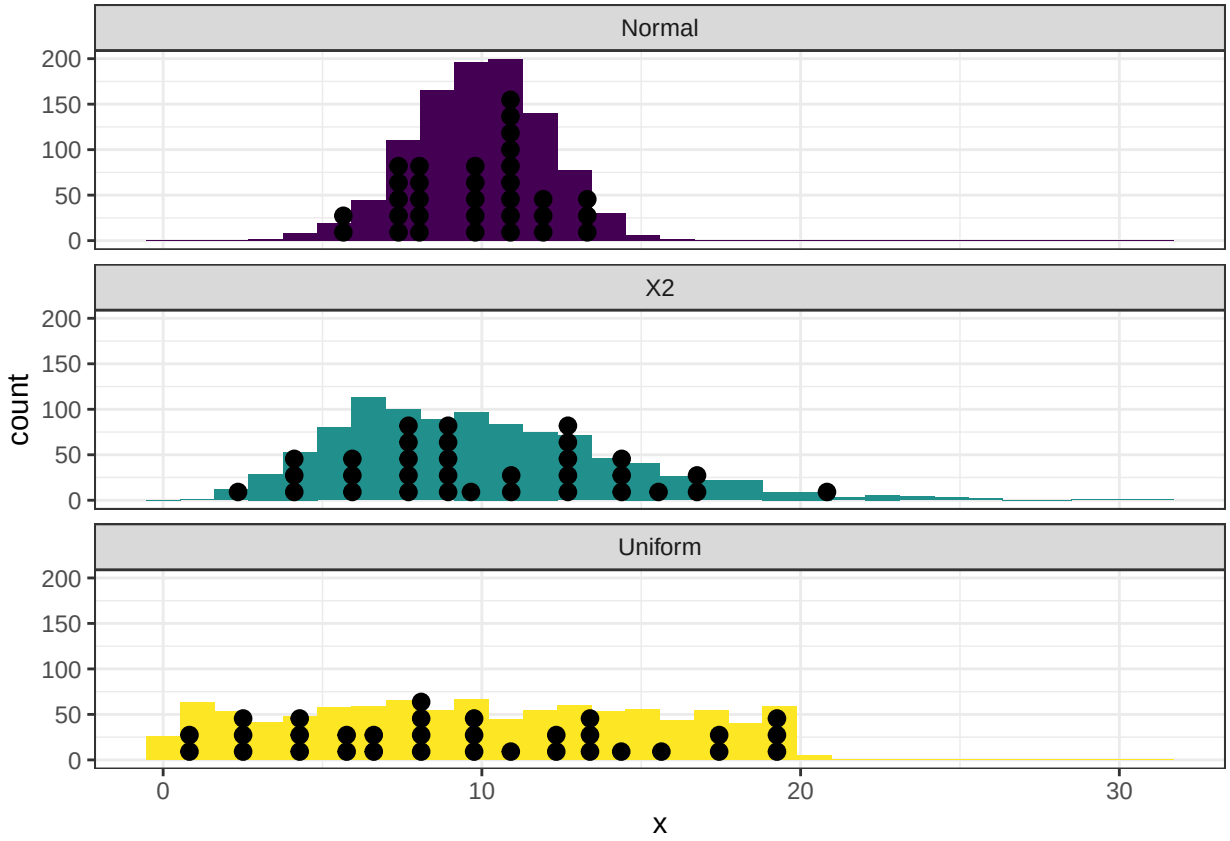
The uniform distribution has a mean of 9.86 and a variance of 32.40. The data ranges from .04 to 20.0 and resembles a uniform distribution.

2. Set p=20% of values missing completely at random (MCAR), then compare the observed distribution to the population.

Set values to missing

```
# pull sample of n=40
s=40
set.seed(8675309)
norm.samp <- sample(norm.pop.data$x, size=s, replace=TRUE)
chi2.samp <- sample(chi2.pop.data$x, size=s, replace=TRUE)
unif.samp <- sample(unif.pop.data$x, size=s, replace=TRUE)

# Create missing (make sure to match the distribution names exactly)
obs.norm <- data.frame(x=create.MCAR(dta=norm.samp, pmiss=.2), dist="Normal")
obs.chi2 <- data.frame(x=create.MCAR(dta=chi2.samp, pmiss=.2), dist="X2")
obs.unif <- data.frame(x=create.MCAR(dta=unif.samp, pmiss=.2), dist="Uniform")

# combine into one data set
```

```r
observed <- rbind(obs.norm, obs.chi2, obs.unif)

# plot population distribution, with sample points added on in black.
ggplot(pop.data, aes(x=x, fill=dist)) + geom_histogram()  +
      facet_wrap(~dist, ncol=1) + scale_fill_discrete(guide=FALSE)  +
      geom_dotplot(data=observed, aes(x=x), dotsize=.5, fill="black")
```



The sample distribution resembles the normal population extremely well, however some bins dont have any data in them while the uniform distribution seems slightly less uniform when observing the sample data. The chi sqare sample remains right skewed but the sample data seems slightly more right skewed.

3. How does missing data affect inference? How much bias is there? Does the shape of the distribution matter?

```r
# calculate grouped summary stats & approx 95% CI's
mcar.ci <- observed %>% group_by(dist) %>%
                    summarise(xbar=mean(x),
                              lower = xbar-2*sd(x)/sqrt(n()),
                              upper = xbar+2*sd(x)/sqrt(n()))

# take the true mean and distribution name from the population summary stats
compare.mcar <- pop.ss %>% select(dist, mean) %>%
# then left join on the sample CI from the mcar observed data
                    left_join(mcar.ci) %>%
```

```
# calculate the bias, and a logical indicator for if the CI covers the mean.
                    mutate(abs.bias = xbar-mean,
                           pct.bias = (abs.bias/mean)*100,
                           cover = ((mean < upper) & (mean > lower)))

# pretty display
kable(compare.mcar, digits=4) %>% kable_styling(full_width=FALSE)
```

dist	mean	xbar	lower	upper	abs.bias	pct.bias	cover
Normal	9.9468	9.7340	8.9884	10.4796	-0.2128	-2.1394	TRUE
X2	10.0059	10.1101	8.5722	11.6480	0.1042	1.0412	TRUE
Uniform	9.8624	9.5854	7.6076	11.5631	-0.2770	-2.8087	TRUE

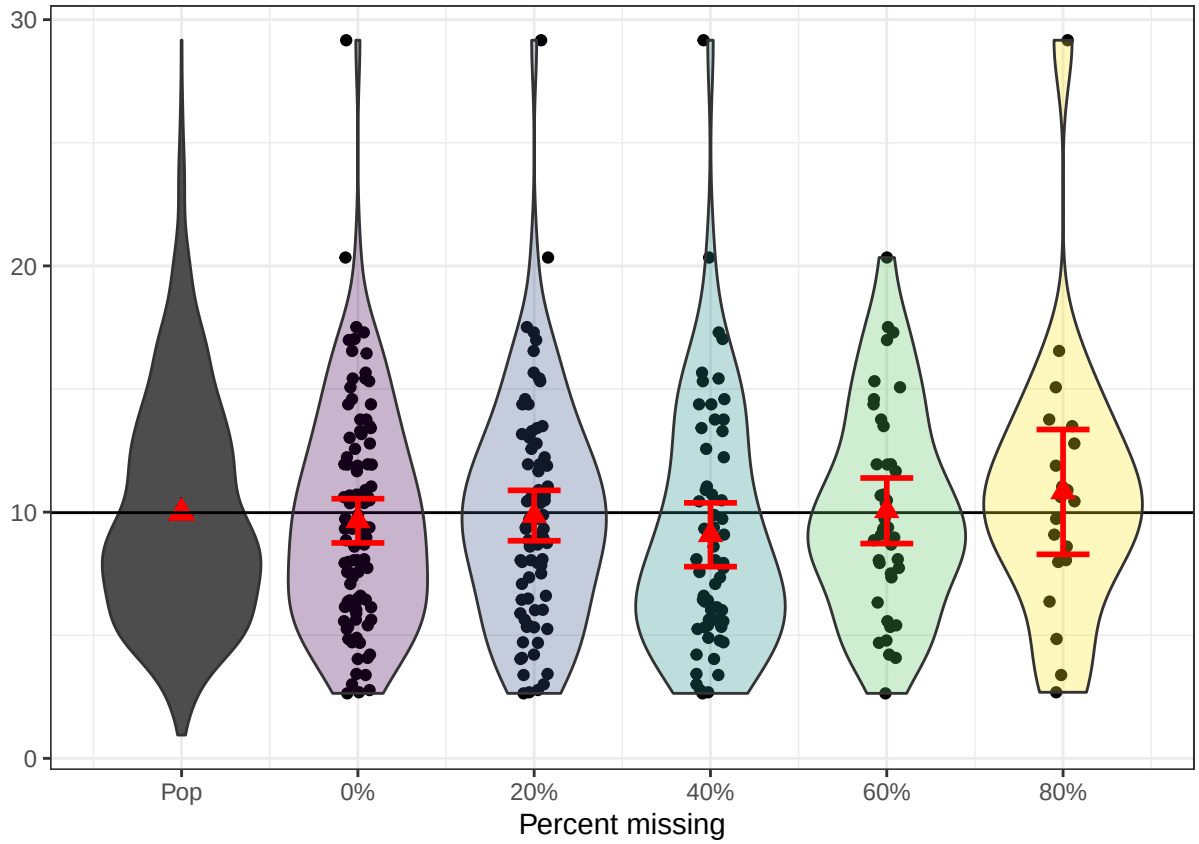
The table above tells us that the confidence interval for the 3 sample distribution cover the population parameter. The means seem to stay fairly accurate when comparing the missing data to the complete data. The percentage bias was .75% for the normal distribution, .64% for the chi sqare, and 4.1% for the uniform which are all relatively low. In addition, for all 3 cases the spread of the data when it is changed to missing is drastically more narrow.

4. How does the % missing change the above results?

```
N=1000
s=100
set.seed(1067)

chi2.pop <- data.frame(x=rchisq(N,10), dist = "X2")
mu = mean(chi2.pop$x)
chi2.0   <- data.frame(x=sample(chi2.pop$x, size=s, replace=TRUE), dist="0%")

# Create missing
chi2.20 <- data.frame(x=create.MCAR(dta=chi2.0$x, pmiss=.2), dist="20%")
chi2.40 <- data.frame(x=create.MCAR(dta=chi2.0$x, pmiss=.4), dist="40%")
chi2.60 <- data.frame(x=create.MCAR(dta=chi2.0$x, pmiss=.6), dist="60%")
chi2.80 <- data.frame(x=create.MCAR(dta=chi2.0$x, pmiss=.8), dist="80%")

chi2.obs <- rbind(chi2.0, chi2.20, chi2.40, chi2.60, chi2.80)

chi2.mcar.ci <- chi2.obs %>% group_by(dist) %>%
                             summarise(xbar=mean(x),
                             lower = xbar-2*sd(x)/sqrt(n()),
                             upper = xbar+2*sd(x)/sqrt(n()))

chi2.mcar.ci %>% mutate(abs.bias = xbar-mu,
                        pct.bias = (abs.bias/mu)*100,
                        cover = ((mu < upper) & (mu > lower))) %>%
  kable(digits=4) %>% kable_styling(full_width=FALSE)
```

dist	xbar	lower	upper	abs.bias	pct.bias	cover
0%	9.6470	8.7494	10.5446	-0.3310	-3.3176	TRUE
20%	9.8607	8.8372	10.8841	-0.1174	-1.1762	TRUE
40%	9.0809	7.7854	10.3764	-0.8971	-8.9908	TRUE
60%	10.0549	8.7235	11.3862	0.0769	0.7702	TRUE
80%	10.8206	8.2859	13.3553	0.8426	8.4442	TRUE

```
ggplot(chi2.pop, aes(y=x, x=0)) +
        # pop stuff first
        geom_violin(fill="grey30") +
        geom_point(aes(x=0, y=mu), pch=17, size=3, col="red") +
        geom_hline(yintercept=mu) +
        # now samples
        geom_jitter(data=chi2.obs, aes(y=x, x=as.numeric(dist)), width=.08) +
        geom_violin(data=chi2.obs, aes(y=x, x=as.numeric(dist), fill=dist), alpha=.3) +
        stat_summary(data=chi2.obs, aes(group=dist, x=as.numeric(dist)),
                     fun.y = 'mean', geom='point', pch=17, size=3, col="red") +
        geom_errorbar(data=chi2.obs, aes(y=x, x=as.numeric(dist)),
                      stat="summary", fun.data="mean_se", col="red",
                      fun.args = list(mult = 2), width=0.3, size=1) +
        scale_x_continuous(breaks=c(0:5), labels=c("Pop", levels(chi2.obs$dist)))  +
        scale_color_viridis_d(guide=FALSE) + scale_fill_viridis_d(guide=FALSE) +
        xlab("Percent missing") + ylab("")
```

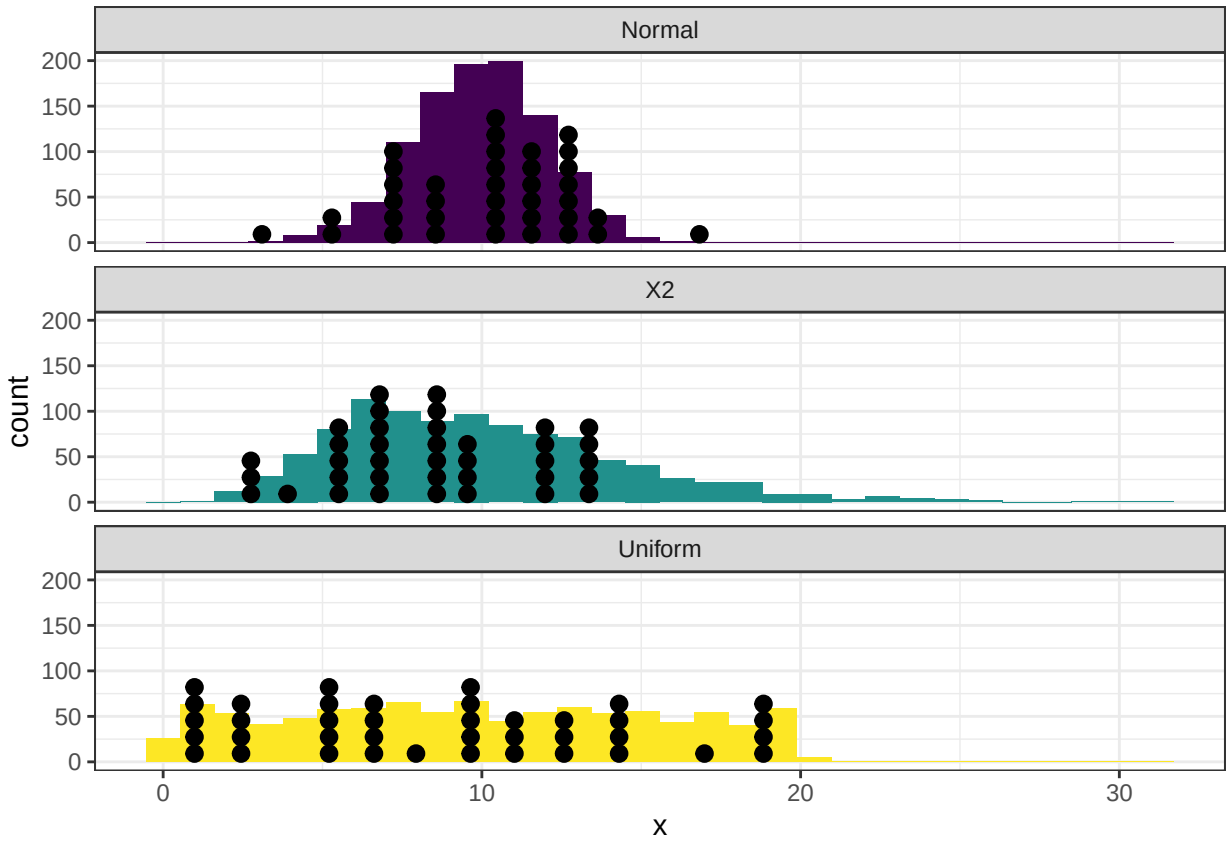


The graph shows that as more data is set to missing, the size of the confidence interval grows slightly. In all cases, the confidence interval contains the mean and it seems that the sample mean is pretty resistant

to change when there is missing data. The table shows us that Bias actually decreases as the missing data changes from 0% to 20% and 20% to 40%. However, after 40% the bias starts to increase substantially goint from 4.83 to 6.88 for 60% missing and finally 17.24 for 80% missing.

5. Repeat problems 2 and 3 using a different missing data mechanism.

```
#Sets variables and seed
s=40
set.seed(8675309)

# sample data from known/named distributions.
norm.pop.data <- data.frame(x=rnorm(n=N, mean=10, sd=2), dist = "Normal")
chi2.pop.data <- data.frame(x=rchisq(N,10), dist = "X2")
unif.pop.data <- data.frame(x=runif(N, 0,20 ), dist = "Uniform")

norm.samp <- sample(norm.pop.data$x, size=s, replace=TRUE)
chi2.samp <- sample(chi2.pop.data$x, size=s, replace=TRUE)
unif.samp <- sample(unif.pop.data$x, size=s, replace=TRUE)

# combine into one data set
pop.data <- rbind(norm.pop.data, chi2.pop.data, unif.pop.data)

# calculate grouped summary stats
pop.ss <- pop.data %>% group_by(dist) %>%
                    summarise(mean=mean(x), var=var(x),
                              min=min(x), max=max(x))
N = 3000
# Create missing (make sure to match the distribution names exactly)
obs.norm <- data.frame(x=create.NMAR(dta=norm.samp, pmiss=.2), dist="Normal")
obs.chi2 <- data.frame(x=create.NMAR(dta=chi2.samp, pmiss=.2), dist="X2")
obs.unif <- data.frame(x=create.NMAR(dta=unif.samp, pmiss=.2), dist="Uniform")

# combine into one data set
observed <- rbind(obs.norm, obs.chi2, obs.unif)

# plot population distribution, with sample points added on in black.
ggplot(pop.data, aes(x=x, fill=dist)) + geom_histogram()  +
      facet_wrap(~dist, ncol=1) + scale_fill_discrete(guide=FALSE)  +
      geom_dotplot(data=observed, aes(x=x), dotsize=.5, fill="black")
```




```
# calculate grouped summary stats & approx 95% CI's
nmar.ci <- observed %>% group_by(dist) %>%
                    summarise(xbar=mean(x),
                              lower = xbar-2*sd(x)/sqrt(n()),
                              upper = xbar+2*sd(x)/sqrt(n()))

# take the true mean and distribution name from the population summary stats
compare.nmar <- pop.ss %>% select(dist, mean) %>%
# then left join on the sample CI from the mcar observed data
                    left_join(nmar.ci) %>%
# calculate the bias, and a logical indicator for if the CI covers the mean.
                    mutate(abs.bias = xbar-mean,
                           pct.bias = (abs.bias/mean)*100,
                           cover = ((mean < upper) & (mean > lower)))

# pretty display
kable(compare.nmar, digits=4) %>% kable_styling(full_width=FALSE)
```

dist	mean	xbar	lower	upper	abs.bias	pct.bias	cover
Normal	9.9468	NA	NA	NA	NA	NA	NA
X2	10.0059	NA	NA	NA	NA	NA	NA
Uniform	9.8624	NA	NA	NA	NA	NA	NA

Using the not missing at random method, we can see that the sample data for the normal distribution did not pick many values at the tails of the graph. We can also see that the chi square distribution is slightly more

right skew to it. However, the uniform distribution's sample data does not look as uniform as the population. There are more data values towards the ends of the graph and less in the middle of the distribution. Looking at our table, all of our summary statistics are showing up as NA.

6. What do you think would happen if the missing data mechanism was negatively correlated with the value of x?

If the missing data mechanism was negatively correlated with x, then the missing data itself could be described by another missing data point.

Multiple Imputation using Chained Equations

1. What percent of the data set overall is missing?

```
hiv.all <- readRDS("C:/Users/KBari/OneDrive/Desktop/Math 456/Hiv_data.rds")
hiv <- hiv.all %>%
       select(age, gender, bsi_overall, frnds, hookey, likesch, school,
              jobmo, edumo, howrel, attserv, livwith, finsit)

survey <- MASS::survey

round(prop.table(table(is.na(hiv)))*100,1)
```

```
##
## FALSE  TRUE
##  94.4   5.6
```

Overall 5.6% of the variables came back as missing.

2. How much missing data is there per variable?

```
prop.miss <- apply(hiv, 2, function(x) round(sum(is.na(x))/NROW(x),4))
prop.miss
```

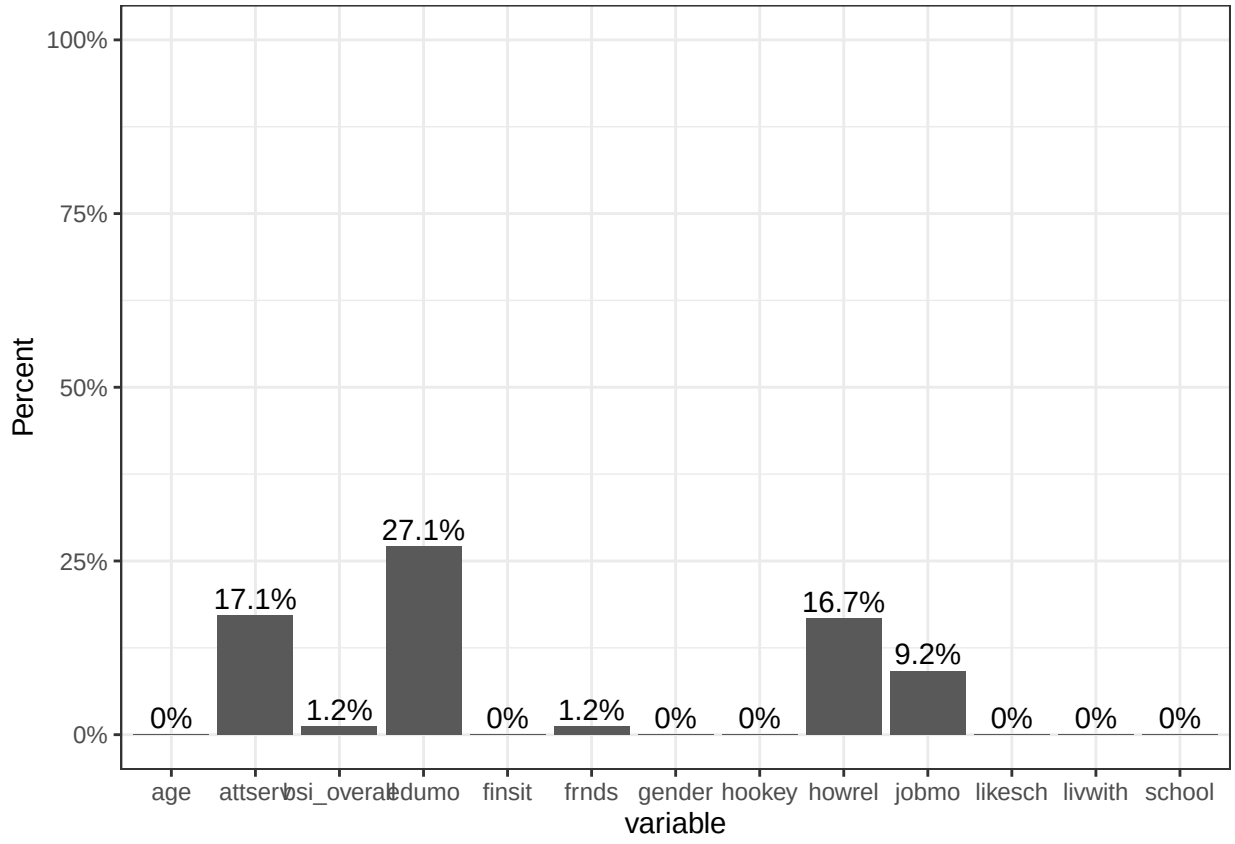
```
##         age      gender bsi_overall       frnds      hookey     likesch
##      0.0000      0.0000      0.0120      0.0120      0.0000      0.0000
##      school       jobmo       edumo      howrel     attserv     livwith
##      0.0000      0.0916      0.2709      0.1673      0.1713      0.0000
##      finsit
##      0.0000
```

27.1% of the responses for mother's education are missing. 17.1% of the responses for how often they attend religious services are missing. 16.7% of the responses for how religious the individual was are missing. 9.2% of the responses for job of the mother are missing. 1.2% of responses for number of friends as well as BSI are missing.

All other variables are less than 1% missing for this set.

3. Describe the missing data patterns.

```
pmpv <- data.frame(variable = names(hiv), pct.miss =prop.miss)

ggplot(pmpv, aes(x=variable, y=pct.miss)) +
  geom_bar(stat="identity") + ylab("Percent") + scale_y_continuous(labels=scales::percent, limits=c(0,1
  geom_text(data=pmpv, aes(label=paste0(round(pct.miss*100,1),"%"), y=pct.miss+.025), size=4)
```



I believe the questions about religion would be considered not missing at random because people might not want to comment if they do not go to religious services frequently or practice a certain religion. For missing data regarding the job and education of the mother, i checked to see how many respondants still live with their mother

```
round(prop.table(table(hiv$livwith))*100, 1)
```

```
##
## Both parents   One parent        Other
##         18.7         71.7          9.6
```

The amount of people that no longer live with their mother is at least 9.6%, this could describe why 9.2% of the data for jobmo is missing.

4a. Use tables to examine the relationship in missing data between 1) the two religion variables `howrel` and `attserv`, 2) between the two financial variables `jobmo` and `finsit`, and 3) how much they like school `likesch` and if they have played hookey `hookey`.

```
pander(table(hiv$howrel,hiv$attserv, useNA = "always"))
```

	Frequently	Never	Sometimes	NA
Not at all	4	28	17	0
Somewhat	41	45	44	2
Very	17	4	7	0
NA	0	1	0	41

```
pander(table(hiv$jobmo,hiv$finsit, useNA = "always"))
```

	Comfortable	Necessities	Poor	Very poor	NA
Employed	16	11	2	0	0
Retired/Disabled	45	19	11	7	0
Unemployed	70	31	12	4	0
NA	12	7	3	1	0

```
pander(table(hiv$likesch,hiv$hookey, useNA = "always"))
```

	0	1	NA
dislike somewhat	3	12	0
dislike very much	9	24	0
meh	28	39	0
somewhat	32	45	0
Very much	31	28	0
NA	0	0	0

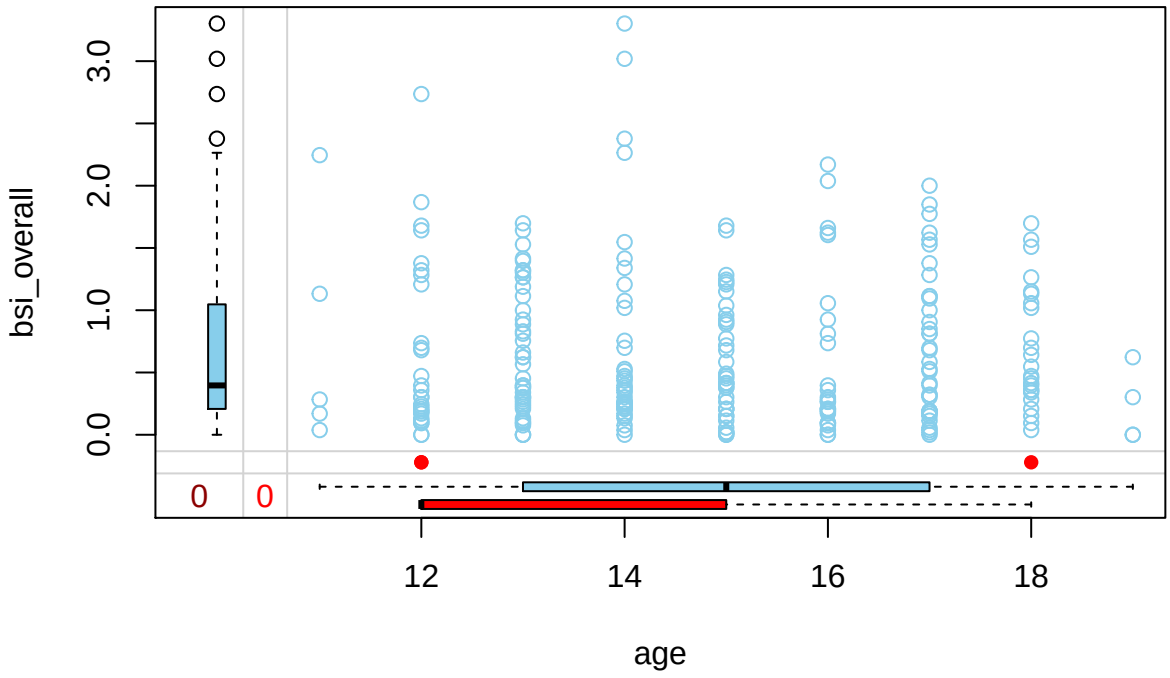
Looking at the variables for how religious a person is and how often they attend service, we can see that these variables are usually missing together. There were 41 instances where both of these were missing together.

Looking at the job of the mother and financial situation, we can see that there is no data missing from financial situation. Job of the mother seems to be missing more as their financial situation gets better however this could be due to the fact that a higher quantity of people with comfortable financial situations were polled.

When we look at the table for if they liked school and if they played hookey, we can see there is no data missing from these.

4b. Use a margin plot to describe the relationship of missing data between Age and BSI overall.

```
View(hiv)
marginplot(hiv[,c(1,3)])
```



There are no missing data points for the age variable. The mean for the missing data seems to be slightly lower than the true mean age.

5. Multiply impute this data set $m = 5$ times. State the imputation models used for each variable.

The imputation models used for each variable are

```
imp_hiv <- mice(hiv, m=5, maxit=25, meth = hiv$meth, seed=500, printFlag=FALSE)
summary(imp_hiv)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##         age      gender bsi_overall       frnds      hookey     likesch
##          ""          ""       "pmm"   "polyreg"          ""          ""
##      school       jobmo       edumo      howrel     attserv     livwith
##          ""   "polyreg"   "polyreg"   "polyreg"   "polyreg"          ""
##      finsit
##          ""
## PredictorMatrix:
```

```
##             age gender bsi_overall frnds hookey likesch school jobmo edumo
## age           0      1           1     1      1       1      1     1     1
## gender        1      0           1     1      1       1      1     1     1
## bsi_overall   1      1           0     1      1       1      1     1     1
## frnds         1      1           1     0      1       1      1     1     1
## hookey        1      1           1     1      0       1      1     1     1
## likesch       1      1           1     1      1       0      1     1     1
##             howrel attserv livwith finsit
## age              1       1       1      1
## gender           1       1       1      1
## bsi_overall      1       1       1      1
## frnds            1       1       1      1
## hookey           1       1       1      1
## likesch          1       1       1      1
```

6. After controlling for other measures, what is the effect of gender on the likelihood a student will skip school? Adjust the model for fit or stability as needed. Report your results in a nice table and interpret the effect of gender on skipping school.

```
complete.model <- glm(hookey ~ age + edumo + jobmo + likesch + howrel + gender, data=hiv,family = "bino

imp.model <- with(imp_hiv, glm(hookey ~ age + edumo + jobmo + likesch + howrel + gender, family = "bino
```

```
pander(summary(complete.model))
```

6a. Fit this model on the complete cases (no imputation).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.892	984.6	0.009031	0.9928
age	0.5638	0.1167	4.833	1.347e-06
edumoHS/GED	-0.3636	0.5016	-0.7249	0.4685
edumoHS+	-0.1949	0.5193	-0.3753	0.7075
jobmoRetired/Disabled	0.8954	0.6713	1.334	0.1823
jobmoUnemployed	1.065	0.653	1.631	0.1028
likeschdislike very much	-16.05	984.6	-0.0163	0.987
likeschmeh	-16.74	984.6	-0.017	0.9864
likeschsomewhat	-16.96	984.6	-0.01722	0.9863
likeschVery much	-16.75	984.6	-0.01701	0.9864
howrelSomewhat	-0.5079	0.5518	-0.9204	0.3574
howrelVery	-1.912	0.7467	-2.561	0.01044
genderMale	-0.4165	0.4445	-0.9369	0.3488

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	195.3 on 145 degrees of freedom
Residual deviance:	145.7 on 133 degrees of freedom

The model states that while holding other variables constant, Males are .41 times less likely to play hookey.

6b. Fit this model on the multiply imputed data sets and report the pooled estimates and intervals. Interpret the effect of gender on playing hookey. Did it change from the complete case model?

```
pander(summary(pool(imp.model)))
```

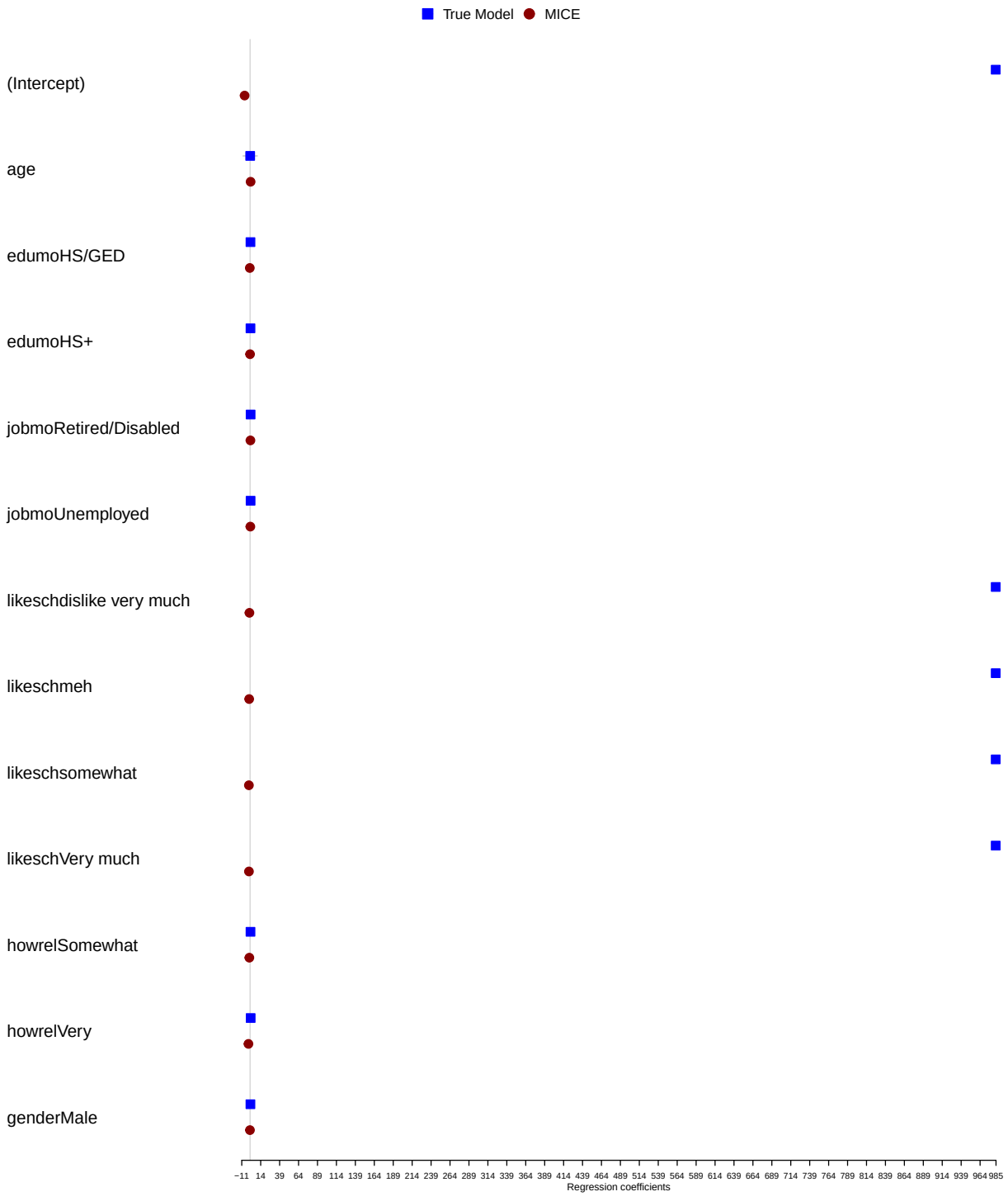
term	estimate	std.error	statistic	df	p.value
(Intercept)	-7.236	1.601	-4.52	224.2	1.002e-05
age	0.6668	0.0995	6.702	212.6	1.819e-10
edumoHS/GED	-0.3743	0.4336	-0.8633	68.31	0.391
edumoHS+	-0.05177	0.4419	-0.1172	75.09	0.9071
jobmoRetired/Disabled	0.4452	0.545	0.8168	217.8	0.4149
jobmoUnemployed	0.3329	0.5082	0.655	213.6	0.5132
likeschdislike very much	-0.8697	0.9062	-0.9598	214.1	0.3382
likeschmeh	-1.221	0.8086	-1.51	233.4	0.1325
likeschsomewhat	-1.578	0.811	-1.945	229.7	0.05295
likeschVery much	-1.515	0.8216	-1.844	231.1	0.06653
howrelSomewhat	-0.9885	0.4485	-2.204	124.1	0.02937
howrelVery	-2.175	0.6429	-3.384	133.7	0.000939
genderMale	-0.1918	0.3271	-0.5864	231	0.5582

The imputed model shows us that males are only .23 times less likely than females to play hookey however neither of the models were statistically significant and both had high p values. (p>.30)

7. Create a forestplot to compare the results for all coefficients in the model.

What are the biggest differences you notice? Would the inference/interpretation of the effect of any covariate on the odds of a student skipping school change depending on what model you use?

```
te.mean <- summary(complete.model)$coefficients[,2]
mi.mean <- summary(pool(imp.model))[,2]
te.ll <- te.mean - 1.96*summary(complete.model)$coefficients[,3]
mi.ll <- mi.mean - 1.96*summary(pool(imp.model))[,3]
te.ul <- te.mean + 1.96*summary(complete.model)$coefficients[,3]
mi.ul <- mi.mean + 1.96*summary(pool(imp.model))[,3]
names <- names(coef(complete.model))

forestplot(names,
legend = c("True Model", "MICE"),
fn.ci_norm = c(fpDrawNormalCI, fpDrawCircleCI),
mean = cbind(te.mean, mi.mean),
lower = cbind(te.ll, mi.ll),
upper = cbind(te.ul, mi.ul),
col=fpColors(box=c("blue", "darkred")),
xlab="Regression coefficients",
boxsize = .1
)
```



Whether or not they like school varies between the imputed and original model, hovwever the models for the rest of the variables are similar between the two models.