Randomization, Hypothesis Testing, Bootstrapping Confidence Intervals

Gabe Wallon

2023-03-09

US Regional Mortality Rates Study

Randomization

This data includes the biological sex and age-adjusted mortality rate per 100,000 people for a variety of causes. The goal will be to determine if the mean mortality rates by `sex` (male or female) and by `status` (living in Urban or Rural area) for the sample differ significantly between the populations in question.

```
data("USRegionalMortality")

sex_summary <- USRegionalMortality %>%
  group_by(Sex) %>%
  summarize(Count=n(),
            AvgRate=mean(Rate)) %>%
  ungroup()
sex_stat = sex_summary$AvgRate[2] - sex_summary$AvgRate[1]

status_summary <- USRegionalMortality %>%
  group_by(Status) %>%
  summarize(Count=n(),
            AvgRate=mean(Rate)) %>%
  ungroup()
status_stat = status_summary$AvgRate[1] - status_summary$AvgRate[2]
```

This sample indicates that, in general, males in the US tend to have higher mortality rates than females, and that people living in rural areas tend to have higher mortality rates than people living in urban areas. Specifically the sample shows that the mortality rate for men is around 20% higher than that for women, and that the mortality rate for those in rural areas is around 8% larger than that for those in urban areas.

Sex Hypotheses:
H_0: The true mean mortality rate for males (μ_M) is equivalent to the true mean mortality rate for females (μ_F).
H_A: The true mean mortality rate for males (μ_M) is greater than the true mean mortality rate for females (μ_F).

$$H_0 : \mu_M - \mu_F = 0$$
$$H_A : \mu_M - \mu_F > 0$$

Status Hypotheses:
H_0: The true mean mortality rate for those in rural areas (μ_R) is equivalent to the true mean mortality rate for those in urban areas (μ_U).
H_0: The true mean mortality rate for those in rural areas (μ_R) is greater than the true mean mortality rate

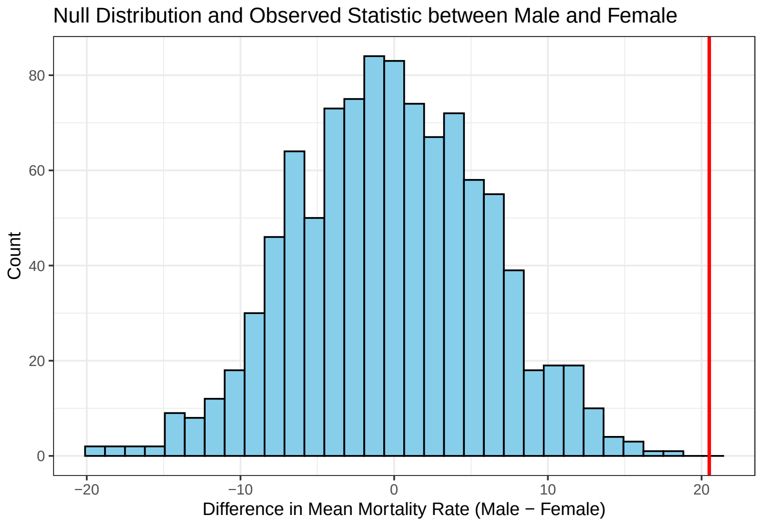
for those in urban areas (μ_U).

$$H_0 : \mu_R - \mu_U = 0$$
$$H_A : \mu_R - \mu_U > 0$$

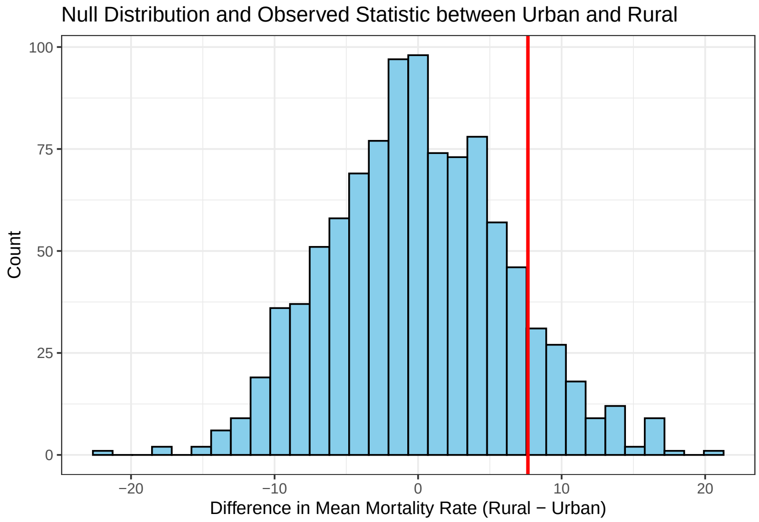
I will choose a level of significance $\alpha = 0.05$ for both of these tests.

Now will generate null distributions, to get an idea of how likely these differences in mortality rates suggested by the sample would be given there were no real difference in mortality rates between these populations.

```r
#initiate empty vector
results_sex <- data.frame(diff=rep(NA,1000))
results_status <- data.frame(diff=rep(NA,1000))

#repeat randomization process 1000 times and save results
for(i in 1:1000){
  # sex
  Shuffle1 <- USRegionalMortality %>%
    transmute(Sex,Rate,
              Shuffle_Rate=sample(USRegionalMortality$Rate,replace=FALSE)) %>%
    group_by(Sex) %>%
    summarize(AvgRate=mean(Shuffle_Rate)) %>%
    ungroup()

  # status
  Shuffle2 <- USRegionalMortality %>%
    transmute(Status,Rate,
              Shuffle_Rate=sample(USRegionalMortality$Rate,replace=FALSE)) %>%
    group_by(Status) %>%
    summarize(AvgRate=mean(Shuffle_Rate)) %>%
    ungroup()

  results_sex$diff[i] <- Shuffle1$AvgRate[2]-Shuffle1$AvgRate[1]  # male rate - female rate
  results_status$diff[i] <- Shuffle2$AvgRate[1]-Shuffle2$AvgRate[2]  # rural rate - urban rate
}

# graphing the distributions
# sex
ggplot(data=results_sex,aes(x=diff)) +
  geom_histogram(color='black',fill='sky blue',bins=32) +
  geom_vline(xintercept=20.50,color='red',size=1) +
  labs(title='Null Distribution and Observed Statistic between Male and Female',
       x='Difference in Mean Mortality Rate (Male - Female)',y='Count') +
  theme_bw()
```



```
#status
ggplot(data=results_status,aes(x=diff)) +
  geom_histogram(color='black',fill='sky blue',bins=32) +
  geom_vline(xintercept=7.65,color='red',size=1) +
  labs(title='Null Distribution and Observed Statistic between Urban and Rural',
       x='Difference in Mean Mortality Rate (Rural - Urban)',y='Count') +
  theme_bw()
```



We can see from these images that given there were no real difference in mortality rate between males and females, the test statistic from sample data would be very unlikely. However, it would *not* be all that unlikely to see the sample difference in mortality rates between urban and rural populations given a true null hypothesis.

I will now quantify a p-value for both of these test statistics.

```
sex_pval <- results_sex %>%
  summarize(pvalue = mean(diff >= sex_stat))

status_pval <- results_status %>%
  summarize(pvalue = mean(diff >= status_stat))
```

These p-values confirm the previously stated theories. Since `sex_pval` is effectively zero, I am compelled to reject the first null hypothesis related to the sex mortality rate difference. However, with the `status_pval` equal to 0.127, I would fail to reject the null hypothesis related to the status mortality rate difference.

Assuming Normality

We will now assume that the conditions from the sampling procedure (`?USRegionalMortality`) meet the requirements of the Central Limit Theorem to assume that the sampling distribution of the difference of means will be normally distributed.

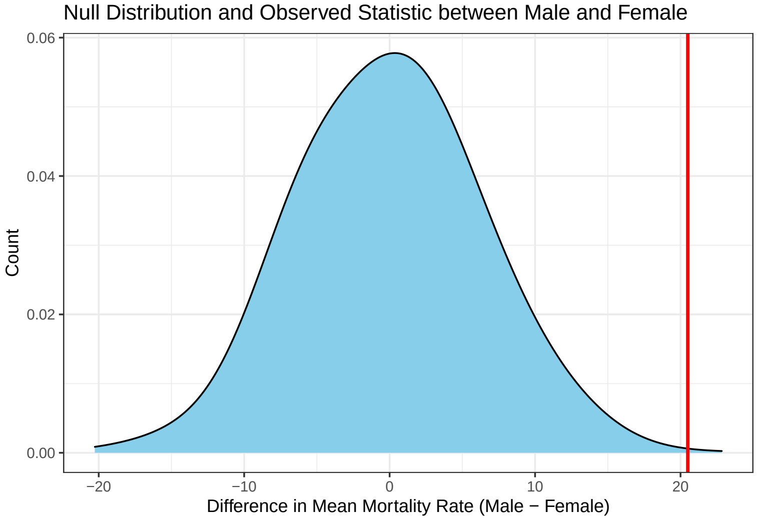
- The observations were collected randomly, thus they are independent from one another.
- There are enough observations from each population (200 from each is likely enough).

Hence, the null distributions in these two cases can be assumed Normal, with the following characteristics:

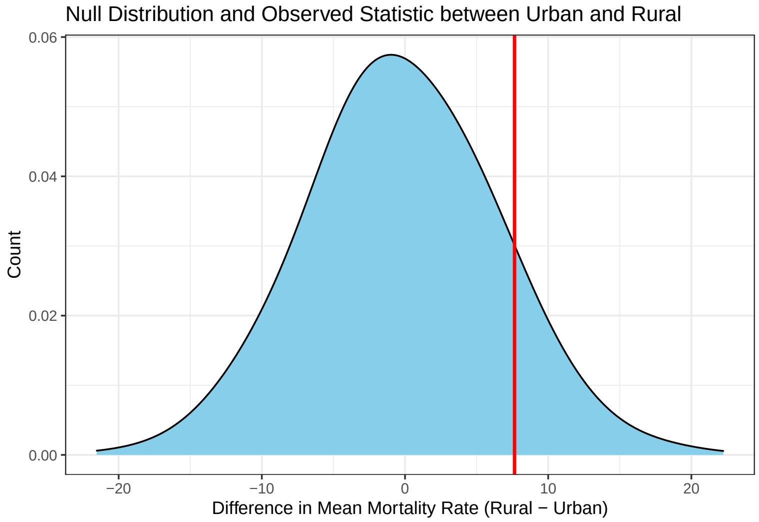
- Center: The difference in true proportions ($\mu_1 - \mu_2$)
- Spread: The true standard error ($\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$)

Where μ_1, μ_2, the respective population means, in both the null hypotheses are assumed to be equal, and the true variances σ_1^2, σ_2^2 will be replaced with the sample variances s_1^2, s_2^2.

```
# sex
sd_male = USRegionalMortality %>% filter(Sex=="Male") %>% summarise(sd(Rate)) %>% pull()
sd_female = USRegionalMortality %>% filter(Sex=="Female") %>% summarise(sd(Rate)) %>% pull()
sterr_sex = sqrt( (sd_male^2 + sd_female^2) / 200 )
sex_diff_norms = data.frame(diffs = rnorm(n=1000,
                                          mean=0,
                                          sd=sterr_sex))

ggplot(data=sex_diff_norms,aes(x=diffs)) +
  geom_density(fill='sky blue',adjust=2) +
  geom_vline(xintercept=20.50,color='red',size=1) +
  labs(title='Null Distribution and Observed Statistic between Male and Female',
       x='Difference in Mean Mortality Rate (Male - Female)',y='Count') +
  theme_bw()
```



```
#status
sd_urban = USRegionalMortality %>% filter(Status=="Urban") %>% summarise(sd(Rate)) %>% pull()
sd_rural = USRegionalMortality %>% filter(Status=="Rural") %>% summarise(sd(Rate)) %>% pull()
sterr_status = sqrt( (sd_urban^2 + sd_rural^2) / 200 )
status_diff_norms = data.frame(diffs = rnorm(n=1000,
                                          mean=0,
                                          sd=sterr_status))

ggplot(data=status_diff_norms,aes(x=diffs)) +
  geom_density(fill='sky blue',adjust=2) +
  geom_vline(xintercept=7.65,color='red',size=1) +
  labs(title='Null Distribution and Observed Statistic between Urban and Rural',
       x='Difference in Mean Mortality Rate (Rural - Urban)',y='Count') +
  theme_bw()
```



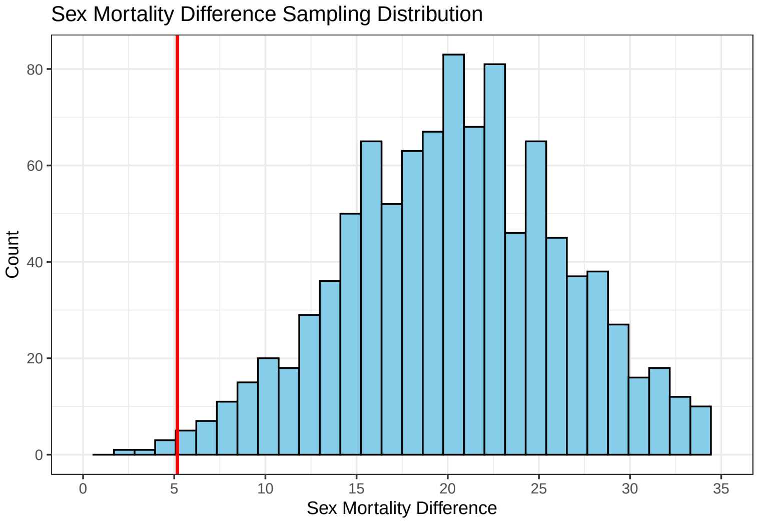
We can tell from these plots that we will get similar p-values and draw the same conclusions as stated previously.

Confidence Intervals

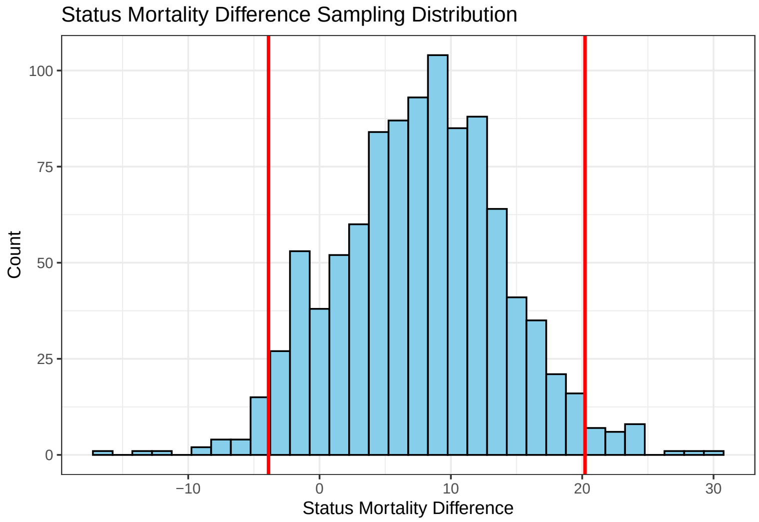
I would like to find a range of plausible values for these two test statistics of interest, thus I will bootstrap confidence intervals for these values.

```
# wrangle data
male_rates <- USRegionalMortality %>%
  select(Sex, Rate) %>%
  filter(Sex=="Male")

female_rates <- USRegionalMortality %>%
  select(Sex, Rate) %>%
  filter(Sex=="Female")

rural_rates <- USRegionalMortality %>%
  select(Status, Rate) %>%
  filter(Status=="Rural")

urban_rates <- USRegionalMortality %>%
  select(Status, Rate) %>%
  filter(Status=="Urban")


# initialize an empty vector
results <- data.frame(sex_diff = rep(NA,1000), status_diff = rep(NA,1000))
```

```
# fill with resampled differences
for(i in 1:1000){
  # sex
  male = mean(sample(male_rates$Rate, size=200, replace=TRUE))
  female = mean(sample(female_rates$Rate, size=200, replace=TRUE))
  results$sex_diff[i] = male - female

  # status
  rural = mean(sample(rural_rates$Rate, size=200, replace=TRUE))
  urban = mean(sample(urban_rates$Rate, size=200, replace=TRUE))
  results$status_diff[i] = rural - urban
}

# graph these plausible ranges of differences
# sex with 99% confidence interval
ggplot(data=results, aes(x=sex_diff)) +
  geom_histogram(color='black',fill='sky blue',bins=32) +
  geom_vline(xintercept = quantile(results$sex_diff, .005),color='red',size=1) +
  geom_vline(xintercept = quantile(results$sex_diff, .995),color='red',size=1) +
  labs(title='Sex Mortality Difference Sampling Distribution',
       x='Sex Mortality Difference',y='Count') +
  scale_x_continuous(limits=c(0,35),breaks=seq(0,35,5)) +
  theme_bw()
```



```
# status with 95% confidence interval
ggplot(data=results, aes(x=status_diff)) +
  geom_histogram(color='black',fill='sky blue',bins=32) +
  geom_vline(xintercept = quantile(results$status_diff, .025),color='red',size=1) +
  geom_vline(xintercept = quantile(results$status_diff, .975),color='red',size=1) +
  labs(title='Status Mortality Difference Sampling Distribution',
       x='Status Mortality Difference',y='Count') +
  # scale_x_continuous(limits=c(0,35),breaks=seq(0,35,5)) +
  theme_bw()
```



These images show that we can be more than 99% confident that the true difference in male and female mortality rate is indeed positive. However we cannot be confident that the rural mortality rate is greater than the urban.