

# Randomization, Hypothesis Testing, Bootstrapping Confidence Intervals

Gabe Wallon

2023-03-09

## US Regional Mortality Rates

### 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 ( $\mu_M$ ) is equivalent to the true mean mortality rate for females ( $\mu_F$ ).

$H_A$ : The true mean mortality rate for males ( $\mu_M$ ) is greater than the true mean mortality rate for females ( $\mu_F$ ).

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

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

Status Hypotheses:

$H_0$ : The true mean mortality rate for those in rural areas ( $\mu_R$ ) is equivalent to the true mean mortality rate for those in urban areas ( $\mu_U$ ).

$H_0$ : The true mean mortality rate for those in rural areas ( $\mu_R$ ) is greater than the true mean mortality rate

for those in urban areas ( $\mu_U$ ).

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

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

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.

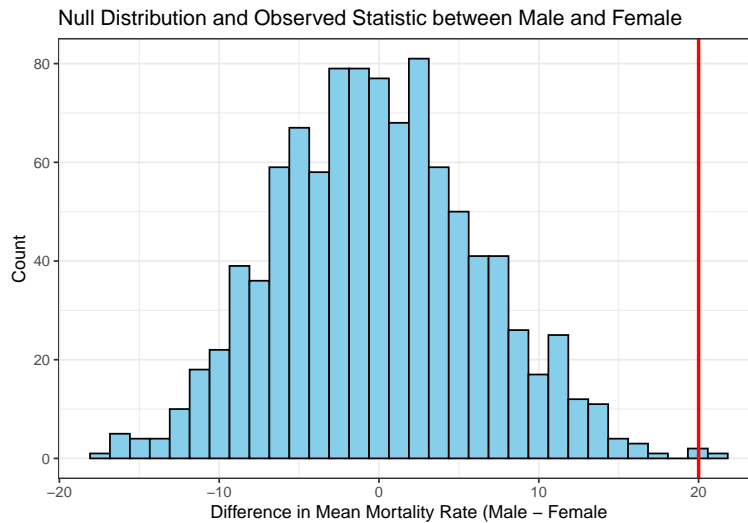
```
#initiate empty vector
results_sex <- data.frame(diff=rep(NA,1000))
results_status <- data.frame(diff=rep(NA,1000))

#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()

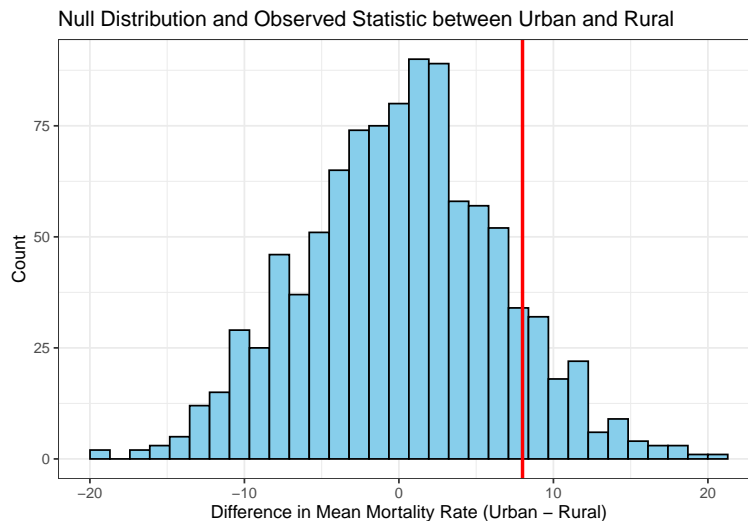
  # status
  Shuffle2 <- USRegionalMortality %>%
    transmute(Status,Rate,
              Shuffle_Rate=sample(USRegionalMortality$Rate,replace=FALSE)) %>%
    group_by(Status) %>%
    summarize(AvgRate=mean(Shuffle_Rate)) %>%
    ungroup()

  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,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=8,color='red',size=1) +
  labs(title='Null Distribution and Observed Statistic between Urban and Rural',
        x='Difference in Mean Mortality Rate (Urban - Rural)',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.

## Confidence Intervals