Appendix

In what follows, we have provided the mathematical explanation of our work.

Statistical Model

Let $P = \{P_\theta : \theta \in \Theta\}$ be a parametrized statistical model and the random variable X represent the suicide rates i.e., the number of suicides per one hundred thousand of the population. We assume that X admits the density P_θ.

The dataset includes suicide rates from 1985 until 2016 which have 27820 rows and 7 columns. The column *'suicide rates'* forms our observations X_i.
Since we are intent on investigating the role of gender in our dataset, we split the dataset based on the *gender* column and formulate two sets of observations for males and females. The samples $X_1^M,...,X_{n_1}^M$ and $X_1^F,...,X_{n_2}^F$ represent males and females' suicide rates, respectively.

Parameters

We denote the mean of the suicide rates for population of males and females by μ_M and μ_F respectively. n_1 and n_2 are the size of the samples of males and females, respectively and both are equal to 13910.

Hypothesis Test

We aim at testing the hypothesis that the mean of population for men is less than mean of population for women. For this purpose, we investigate the quantity $\theta = \mu_M - \mu_F$. For testing our hypothesis, we define our test as following:

$$H_0 : \mu_M \leq \mu_F \qquad \text{against} \qquad H_1 : \mu_M > \mu_F$$

Using notation of θ, the test can also be written as

$$H_0 : \theta \leq 0 \qquad \text{against} \qquad H_1 : \theta > 0$$

Consequently, we have $\Theta_0 = \{\theta : \theta \leq 0\ \}$, $\Theta_1 = \{\theta : \theta > 0\ \}$

Remarks

Since σ for each sample is unknown, we estimate it using the sample variance of two samples which are defined as following

$$S_M^2 = \tfrac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i^M - \overline{X}^M)^2 \qquad S_F^2 = \tfrac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_i^F - \overline{X}^F)^2$$

And the total sample variance is

$$S_T^2 = \tfrac{S_M^2}{n_1} + \tfrac{S_F^2}{n_2}$$

Using central limit theorem, we can show that if $n \to \infty$, then

$$\overline{X}^M - \overline{X}^F \xrightarrow{\text{L}} \mathcal{N}(\mu_M - \mu_F, \sigma_M^2 + \sigma_F^2)\,.$$

Therefore, by the definition of student t's distribution, we have

$$\frac{\left(\overline{X}^M - \overline{X}^F - (\mu_M - \mu_F)\right)}{S_T} \xrightarrow{\text{L}} t_{df=n_1+n_2-2}$$

Estimator

As we are interested in the value $\theta = \mu_M - \mu_F$, we define an estimator of it by

$$\overline{X}^M - \overline{X}^F$$

We claim that this is a consistent estimator of θ. To prove consistency, we can say that by law of large numbers, if n is large enough, then

$$\overline{X}^M - \overline{X}^F \xrightarrow{a.s.} \mathbb{E}[X_1^M] - \mathbb{E}[X_1^F] = \mu_M - \mu_F$$

Test Statistic

In order to accept or reject the test, we define a test statistic as following

$$T(X) = \overline{X}^M - \overline{X}^F$$

Intuitive Approach: finding the threshold

We will use the following quantity as our pivot.

$$\tau = \tfrac{\overline{X}^M - \overline{X}^F - (\mu_M - \mu_F)}{S_T} \sim t_{df=n_1+n_2-2}$$

where degrees of freedom is $n_1 + n_2 - 2$.
For $\theta \in \Theta_0$ or equivalently $\theta \leq 0$,

$$\begin{aligned}
\alpha &= \mathbb{P}(T(X) > u) \\
&= \mathbb{P}(\overline{X}^M - \overline{X}^F > u) \\
&= \mathbb{P}(\frac{\overline{X}^M - \overline{X}^F - (\mu_M - \mu_F)}{S_T} > \frac{u - (\mu_M - \mu_F)}{S_T}) \\
&= \mathbb{P}(\frac{\overline{X}^M - \overline{X}^F - \theta}{S_T} > \frac{u - \theta}{S_T}) \\
&< \mathbb{P}(\frac{\overline{X}^M - \overline{X}^F - \theta}{S_T} > \frac{u}{S_T}) \\
&= \mathbb{P}(\tau > \frac{u}{S_T}) \\
&= \mathbb{P}(\tau > t_\alpha)
\end{aligned} \tag{1}$$

where $t_\alpha = \frac{u}{S_T}$ is the $\{1-\alpha\}$ quantile of the t-distribution and τ is the pivot. The last inequality comes from the fact that since $\theta < 0$, we have $u - \theta > u$.
Solving for threshold u gives that $u = S_T t_\alpha$.

Then the test function ϕ will be

$$\phi(X) = \mathbb{1}_{\{T(X)>u\}} \tag{2}$$

Substituting u into test function will give us the decision

$$\phi(X) = \mathbb{1}_{\{T(X)>S_T t_\alpha\}} \tag{3}$$

Power of Test

The power of test is expressed in the following

$$\begin{aligned}
\mathbb{P}(rejectH_0) &= \mathbb{P}(\Phi(X) = 1) = \mathbb{E}[\Phi] \\
&= \mathbb{P}(\overline{X}^M - \overline{X}^F > u) \\
&= \mathbb{P}(\frac{\overline{X}^M - \overline{X}^F - (\mu_M - \mu_F)}{S_T} > \frac{u - (\mu_M - \mu_F)}{S_T}) \\
&= \mathbb{P}(\frac{\overline{X}^M - \overline{X}^F - \theta}{S_T} > \frac{u - \theta}{S_T}) \\
&= 1 - \phi(\frac{u-\theta}{S_T})
\end{aligned} \tag{4}$$

where ϕ is the cdf of the t-distribution.
For $\theta \in \Theta_1$ or equivalently $\theta > 0$, since $\theta > 0$, $u - \theta$ is a decreasing function of θ. As ϕ is an increasing function, $\phi(\frac{u-\theta}{S_T})$ is also decreasing. Therefore, $1 - \phi(\frac{u-\theta}{S_T})$ will be a increasing function of θ.

Results

$$S_M = 23.55, \qquad S_F = 7.35, \qquad S_T = 0.21,$$

$$T(X) = 14.85, \qquad t_{alpha} = 1.65, \qquad S_T = 0.21, \qquad u = 0.34$$

Since $T(x) > u$, we reject the null hypothesis and we increase our confidence in H_1, which states that the mean for males is greater that mean for females.