Solution to HW question $\mathbb{V}(\mathbf{T})$

Yotam Shem-Tov

Fall 2014

The difference in means

- One of the most common test statistics is the difference in means,
$$T = \frac{\sum_{i=1}^{N} Z_i Y_i}{m} - \frac{\sum_{i=1}^{N}(1 - Z_i) Y_i}{N - m}$$
- Denote the outcomes with treatment as a_i and with control as b_i
- Consider the null hypothesis that the treatment has no effect, i.e $a_i = b_i$
- Under the null hypothesis what is the expectation of difference in means estimator,
$$\mathbb{E}\{T\} = \mathbb{E}\{\bar{a} - \bar{b}\} = \mathbb{E}\{a_i\} - \mathbb{E}\{b_i\} = 0$$
- What is the variance of the difference in means estimator? $\mathbb{V}(T) =?$

The difference in means

- One of the most common test statistics is the difference in means,
$$T = \frac{\sum_{i=1}^{N} Z_i Y_i}{m} - \frac{\sum_{i=1}^{N}(1 - Z_i) Y_i}{N - m}$$
- Denote the outcomes with treatment as a_i and with control as b_i
- Consider the null hypothesis that the treatment has no effect, i.e $a_i = b_i$
- Under the null hypothesis what is the expectation of difference in means estimator,
$$\mathbb{E}\{T\} = \mathbb{E}\{\bar{a} - \bar{b}\} = \mathbb{E}\{a_i\} - \mathbb{E}\{b_i\} = 0$$
- What is the variance of the difference in means estimator? $\mathbb{V}(T) =$? Homework question - the answer is not as easy as it might seem

The difference in means: variance calculation

- In a finite sample $x_1, \ldots, x_N$, the expectation is

$$\mathbb{E}(x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and the variance is,

$$\mathbb{V}(x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2$$

- What is finite sample correction?

The difference in means: variance calculation

- In a finite sample $x_1, \ldots, x_N$, the expectation is

$$\mathbb{E}(x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and the variance is,

$$\mathbb{V}(x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2$$

- What is finite sample correction?
  In order to adjust the variance for sampling from a finite population we need to adjust the variance by, $\frac{N-n}{N-1}$. Where N is the population size and n is the sample size
- When the population size is not infinite relative to the sample size, we need to use a finite sample correction

The difference in means: variance calculation hints

- What is $\mathbb{V}(\bar{a}) =?$

- What is $\mathbb{V}(\bar{a}) =?$

$$\mathbb{V}(\bar{a}) = \frac{1}{m} \cdot \mathbb{V}(a_i) \left(\frac{N-m}{N-1} \right)$$

$$= \left(\frac{N-m}{N-1} \right) \frac{1}{m} \left(\frac{1}{N} \sum_{i=1}^{N} a_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} a_i \right)^2 \right)$$

- What is $\mathbb{V}(\bar{b}) =?$

The difference in means: variance calculation hints

- What is $\mathbb{V}(\bar{a}) =$?

$$\mathbb{V}(\bar{a}) = \frac{1}{m} \cdot \mathbb{V}(a_i) \left(\frac{N-m}{N-1} \right)$$

$$= \left(\frac{N-m}{N-1} \right) \frac{1}{m} \left(\frac{1}{N} \sum_{i=1}^{N} a_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} a_i \right)^2 \right)$$

- What is $\mathbb{V}(\bar{b}) =$?

$$\mathbb{V}(\bar{b}) = \frac{1}{N-m} \cdot \mathbb{V}(b_i) \left(\frac{N-(N-m)}{N-1} \right)$$

$$= \left(\frac{N-(N-m)}{N-1} \right) \frac{1}{N-m} \left(\frac{1}{N} \sum_{i=1}^{N} b_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} b_i \right)^2 \right)$$

Solution to HW question
The difference in means variance: Analytical Solution

$$\mathbb{V}\left(\bar{a} - \bar{b}\right) = \mathbb{V}\left(\bar{a}\right) + \mathbb{V}\left(\bar{a}\right) - 2Cov(\bar{a}, \bar{b})$$

The difference in means variance: Analytical Solution

$$\mathbb{V}\left(\bar{a}-\bar{b}\right)=\mathbb{V}\left(\bar{a}\right)+\mathbb{V}\left(\bar{a}\right)-2Cov(\bar{a},\bar{b})$$

Denote by σ_a^2 the variance in the treatment group, and by σ_b^2 the variance in the control group. Under the null, $\sigma_a^2=\sigma_b^2=\sigma^2$

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}b_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}b_i\right)^2=\frac{1}{N}\sum_{i=1}^{N}a_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}a_i\right)^2$$

Hence,

Solution to HW question
The difference in means variance: Analytical Solution

$$\mathbb{V}\left(\bar{a}-\bar{b}\right) = \mathbb{V}\left(\bar{a}\right) + \mathbb{V}\left(\bar{a}\right) - 2Cov(\bar{a},\bar{b})$$

Denote by σ_a^2 the variance in the treatment group, and by σ_b^2 the variance in the control group. Under the null, $\sigma_a^2 = \sigma_b^2 = \sigma^2$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} b_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} b_i\right)^2 = \frac{1}{N}\sum_{i=1}^{N} a_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} a_i\right)^2$$

Hence,

$$\mathbb{V}\left(\bar{a}\right) = \sigma^2\frac{1}{m}\left(\frac{N-m}{N-1}\right)$$

$$\mathbb{V}\left(\bar{b}\right) = \sigma^2\frac{1}{N-m}\left(\frac{m}{N-1}\right)$$

Solution to HW question
The difference in means variance: Analytical Solution

$$Cov(\bar{a}, \bar{b}) = Cov(\frac{1}{m}\sum_i a_i, \frac{1}{N-m}\sum_j b_j)$$

$$= \frac{1}{m(N-m)} Cov(\sum_i a_i, \sum_j b_j) = \frac{1}{m(N-m)} \sum_i \sum_{j \neq i} Cov(a_i, b_j)$$

Solution to HW question
The difference in means variance: Analytical Solution

$$Cov(\bar{a}, \bar{b}) = Cov(\frac{1}{m}\sum_i a_i, \frac{1}{N-m}\sum_j b_j)$$

$$= \frac{1}{m(N-m)} Cov(\sum_i a_i, \sum_j b_j) = \frac{1}{m(N-m)} \sum_i \sum_{j \neq i} Cov(a_i, b_j)$$

$$= \frac{1}{m(N-m)} m(N-m) Cov(a_i, b_j) = Cov(a_i, b_j)$$

$$Cov(a_i, b_j) = \mathbb{E}\{a_i b_j\} - \mathbb{E}\{b_j\}\mathbb{E}\{b_j\}$$

Note,

The difference in means variance: Analytical Solution

$$Cov(\bar{a}, \bar{b}) = Cov(\frac{1}{m}\sum_i a_i, \frac{1}{N-m}\sum_j b_j)$$

$$= \frac{1}{m(N-m)} Cov(\sum_i a_i, \sum_j b_j) = \frac{1}{m(N-m)} \sum_i \sum_{j \neq i} Cov(a_i, b_j)$$

$$= \frac{1}{m(N-m)} m(N-m) Cov(a_i, b_j) = Cov(a_i, b_j)$$

$$Cov(a_i, b_j) = \mathbb{E}\{a_i b_j\} - \mathbb{E}\{b_j\}\mathbb{E}\{b_j\}$$

Note,

$$\mathbb{E}\{b_j\} = \frac{1}{N-m}\sum_{i=1}^{N} b_i = \frac{1}{N-m}\sum_{i=1}^{N} a_i,\ \mathbb{E}\{a_i\} = \frac{1}{m}\sum_{i=1}^{N} a_i = \frac{1}{m}\sum_{i=1}^{N} b_i$$

Solution to HW question
The difference in means variance: Analytical Solution

$$\mathbb{E}\{a_i b_j\} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} a_i b_j = \frac{1}{N(N-1)} \left(\sum_{i=1}^{N} \sum_{j} a_i b_j - \sum_{i=1}^{N} a_i b_i \right)$$

Solution to HW question
The difference in means variance: Analytical Solution

$$\mathbb{E}\{a_i b_j\} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} a_i b_j = \frac{1}{N(N-1)} \left(\sum_{i=1}^{N} \sum_{j} a_i b_j - \sum_{i=1}^{N} a_i b_i \right)$$

Note,

$$\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j} a_i b_j = \frac{1}{N^2} \sum_{i=1}^{N} a_i \sum_{j} b_j = \frac{1}{N} \sum_{i=1}^{N} a_i \frac{1}{N} \sum_{j} b_j = \mathbb{E}(a_i)\mathbb{E}(b_i)$$

and,

Solution to HW question
The difference in means variance: Analytical Solution

$$\mathbb{E}\{a_i b_j\} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} a_i b_j = \frac{1}{N(N-1)} \left(\sum_{i=1}^{N} \sum_{j} a_i b_j - \sum_{i=1}^{N} a_i b_i \right)$$

Note,

$$\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j} a_i b_j = \frac{1}{N^2} \sum_{i=1}^{N} a_i \sum_{j} b_j = \frac{1}{N} \sum_{i=1}^{N} a_i \frac{1}{N} \sum_{j} b_j = \mathbb{E}(a_i)\mathbb{E}(b_i)$$

and,

$$Cov(a_i, b_i) = \frac{1}{N} \sum_{i=1}^{N} a_i b_i - \mathbb{E}(a_i)\mathbb{E}(b_i)$$

$$\Rightarrow \sum_{i=1}^{N} a_i b_i = N(Cov(a_i, b_i) + \mathbb{E}(a_i)\mathbb{E}(b_i))$$

The difference in means variance: Analytical Solution

Hence,

$$\mathbb{E}\{a_i b_j\} = \frac{1}{N(N-1)}\left[N^2\mathbb{E}(a_i)\mathbb{E}(b_i) - N(Cov(a_i, b_i) + \mathbb{E}(a_i)\mathbb{E}(b_i))\right]$$

The difference in means variance: Analytical Solution

Hence,

$$\mathbb{E}\{a_i b_j\} = \frac{1}{N(N-1)}\left[N^2\mathbb{E}(a_i)\mathbb{E}(b_i) - N(Cov(a_i, b_i) + \mathbb{E}(a_i)\mathbb{E}(b_i))\right]$$

$$= \frac{1}{N(N-1)}\left[N(N-1)\mathbb{E}(a_i)\mathbb{E}(b_i) - NCov(a_i, b_i)\right]$$

$$= \mathbb{E}(a_i)\mathbb{E}(b_i) - \frac{1}{N-1}Cov(a_i, b_i)$$

Therefore,

$$Cov(a_i, b_j) = -\frac{1}{N-1}Cov(a_i, b_i) = -\frac{\sigma^2}{N-1}$$

The variance of the difference in means estimator under the null is,

$$\mathbb{V}(T) = \mathbb{V}\left(\bar{a}\right) + \mathbb{V}\left(\bar{a}\right) - 2Cov(\bar{a}, \bar{b})$$

$$= \sigma^2 \frac{1}{m}\left(\frac{N-m}{N-1}\right) + \sigma^2 \frac{1}{N-m}\left(\frac{m}{N-1}\right) - 2\left(-\frac{\sigma^2}{N-1}\right)$$

$$= \sigma^2 \cdot \frac{N^2}{(N-1)m(N-m)}$$

The difference in means variance: Simulation

- When $N = 12$ and $m = 3$, and

```
> set.seed(12345)
> y = rnorm(N,mean=0,sd=1)
> y
 [1]  0.5855288  0.7094660 -0.1093033 -0.4534972
 0.6058875 -1.8179560  0.6300986 -0.2761841 -0.2841597
[12]  1.8173120
```

- What is the variance of the difference in means estimator?
- In order to answer the question we will approximate the variance using a Monte-Carlo simulation

Solution to HW question
The difference in means variance: Simulation

```
m=3; N=12
set.seed(12345)
y = rnorm(N,mean=0,sd=1)
R = 200000
z = c(rep(1,m),rep(0,N-m))
mean.diff<-mean.t <-mean.c <- rep(999,R)
for (i in c(1:R)){
  z0 = sample(z,N)

  mean.c[i] = mean(y[z0==0])
  mean.t[i] = mean(y[z0==1])
  mean.diff[i] = mean(y[z0==1]) - mean(y[z0==0])
}
cov(mean.c,mean.t)
var(mean.diff)
```

The difference in means variance: Simulation

- The simulation estimation of $\mathbb{V}(\bar{a}, \bar{b})$ is 0.3809463, and the real variance is 0.3814015

Solution to HW question
The difference in means variance: Simulation

- The simulation estimation of $\mathbb{V}(\bar{a}, \bar{b})$ is 0.3809463, and the real variance is 0.3814015
- The $Cov(\bar{a}, \bar{b}) = -0.07151277$, and accounts for 37.5% of the variance of the difference in means

Solution to HW question
The difference in means variance: Simulation

- The simulation estimation of $\mathbb{V}(\bar{a}, \bar{b})$ is 0.3809463, and the real variance is 0.3814015
- The $Cov(\bar{a}, \bar{b}) = -0.07151277$, and accounts for 37.5% of the variance of the difference in means
- **Conclusion: Monte Carlo simulations can help overcome difficult computational problems**

Solution to HW question
The difference in means variance: Asymptotic

What is the asymptotic variance of the difference in means estimator?

Solution to HW question
The difference in means variance: Asymptotic

What is the asymptotic variance of the difference in means estimator? A technical approach:

$$\mathbb{V}(T) = \sigma^2 \cdot \frac{N^2}{(N-1)m(N-m)} \to \frac{\sigma^2}{m}$$

as $N \to \infty$

What is the asymptotic variance of the difference in means estimator? A technical approach:

$$\mathbb{V}(T) = \sigma^2 \cdot \frac{N^2}{(N-1)m(N-m)} \to \frac{\sigma^2}{m}$$

as $N \to \infty$

A more intuitive approach:

When $N \to \infty$, Z_i and Z_j are independent, and hence the variance is,

$$\mathbb{V}(T) = \frac{\sigma^2}{N-m} + \frac{\sigma^2}{m} \to \frac{\sigma^2}{m}$$