# Chapter 1

# Variance of the random walk

Studying variance of estimator is important for the construction of confidence interval and testing of hypothesis.

Let's look on the network graph where nodes have correlated values. Now let's assume that correlation between the nodes depends on the distance between them in the following way: nodes at the distance 1 have correlation $\rho$, at the distance 2 correlation $\rho^2$ and so on. If the distance between nodes $i$ and $j$ is $k$ then $corr(X_i, X_j) = \rho^k$.

First, let's look at the line where nodes are correlated as described above. Then $corr(X_i, X_{i+h}) = \rho^h$. Now let's start to collect the values along the line starting from the first node, $X_1, X_2, ..., X_n$. Then we can count variation of the mean of $X_1, X_2, ..., X_n$.

$$var\left[\frac{X_1, X_2, ..., X_n}{n}\right] = var\left[\bar{X}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} cov(X_i, X_j) =$$

$$= \frac{\sigma^2}{n^2}\left(n + 2(n-1)\rho + 2(n-2)\rho^2 + ... + 2\cdot 2\rho^{n-2} + 2\cdot 1\rho^{n-1}\right) =$$

$$= \frac{\sigma^2}{n^2}\left(n + 2\sum_{i=1}^{n-1}(n-i)\rho^i\right) = \frac{\sigma^2}{n^2}\left(n + 2n\sum_{i=1}^{n-1}\rho^i - 2\sum_{i=1}^{n-1} i\rho^i\right) =$$

$$= \frac{\sigma^2}{n}\left(n + 2n\frac{\rho - \rho^n}{1-\rho} - 2\rho\sum_{i=0}^{n-2}(\rho^{i+1})'\right) =$$

$$= \frac{\sigma^2}{n}\left(n + 2n\frac{\rho - \rho^n}{1-\rho} - 2\rho\left(\frac{\rho - \rho^n}{1-\rho}\right)'\right) =$$

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

$$= \frac{\sigma^2}{n^2}\frac{n - n\rho^2 - 2\rho + 2\rho^{n+1}}{(1-\rho)^2}$$

$$var\left[\bar{X}\right] = \frac{\sigma^2}{n^2}\frac{n - n\rho^2 - 2\rho + 2\rho^{n+1}}{(1-\rho)^2}$$

Let's simplify a bit expression for variance by approximated one.

$$var\left[\bar{X}\right] = \frac{\sigma^2}{n^2}\frac{n - n\rho^2 - 2\rho + 2\rho^{n+1}}{(1-\rho)^2} = \frac{\sigma^2}{n}\frac{n - n\rho^2 - 2\rho + 2\rho^{n+1}}{n(1-\rho)^2} =$$

$$= \frac{\sigma^2}{n}\frac{1 - \rho^2 - 2\rho/n + 2\rho^{n+1}/n}{(1-\rho)^2} \simeq \frac{\sigma^2}{n}\frac{1-\rho^2}{(1-\rho)^2} = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$$

$$var\left[\bar{X}\right] = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$$

Approximation is especially good with big $n$ and $\rho$.

If random variables $X_1, X_2, ..., X_n$ were independent then the variance of $\bar{X}$ would be $var_{ind}[\bar{X}] = \frac{\sigma^2}{n}$.

But we consider random variables $X_1, X_2, ..., X_n$ that are dependent with known correlation and the variance in this case is bigger.

$$var\left[\bar{X}\right] = var_{ind}[\bar{X}]\frac{1+\rho}{1-\rho} = var_{ind}[\bar{X}]\left(1 + \frac{2\rho}{1-\rho}\right) > var_{ind}[\bar{X}]$$

The less is correlation between nodes the closer are variances $var\left[\bar{X}\right]$ and $var_{ind}[\bar{X}]$.

Variance with skipping

Let's look at the variance of the next random variable:

$$\bar{X}^k = \frac{X_1 + X_{1+k} + X_{1+2k} + ... + X_{1+(n-1)k}}{n}$$

So $corr(X_{1+ik}, X_{1+(i+h)k}) = \rho^{kh}$. Now let's introduce new random variable $Y_1, Y_2, ..., Y_n$ such that $Y_1 = X_1, Y_2 = X_{1+k}, ..., Y_n = X_{1+(n-1)k}$ and $r = \rho^k$, $\bar{Y} = \bar{X}^k$. Then $corr(Y_i, Y_{i+h}) = corr(X_{1+(i-1)k}, X_{1+(i+h-1)k}) = \rho^{kh} = r^h$.

To sum up we have random variables $Y_1, Y_2, ..., Y_n$ where $corr(Y_i, Y_{i+h} = \rho^{kh} = r^h$. But we already know that

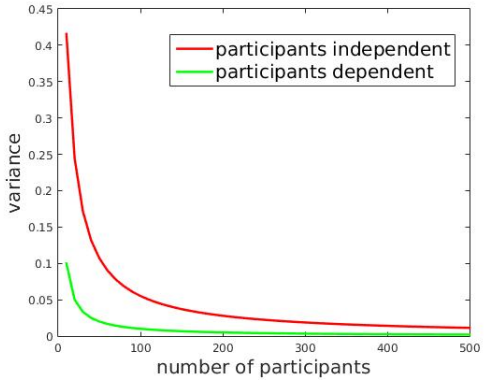


Figure 1.1: $\rho = 0.7$

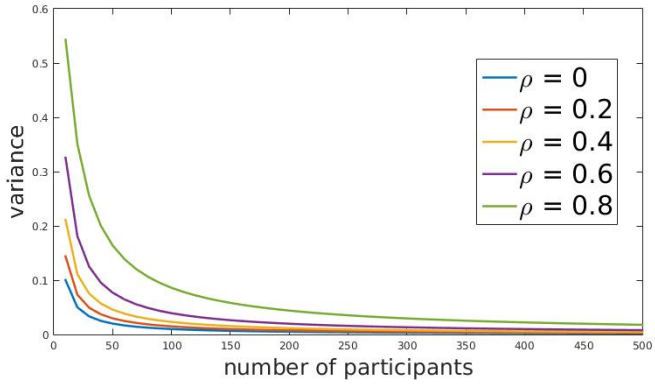


Figure 1.2: $\rho = 0.7$

$$var\left[\bar{Y}\right] \simeq \frac{\sigma^2}{n}\frac{1+r}{1-r}$$

Then

$$var\left[\bar{X}^k\right] \simeq \frac{\sigma^2}{n}\frac{1+\rho^k}{1-\rho^k}.$$

In RDS context

$B$ - budget

$C_1$ - cost of one step of walk (individuals just provide the correct number of their contacts)

$C_2$ - cost of participation (cost of interview with individuals)

$n$ - number of steps
$m$ - number of participants from $n$
The next equality should be true:

$$B = n \cdot C_1 + m \cdot C_2$$

If we want to skip $k$ steps between taking the node as a participant then

$$B = nC_1 + \frac{n}{k+1}C_2$$

Here $m = \frac{n}{k+1}$ as we take each $k+1$ node as a participant. So having budget $B$ and skipping each $k$ node allows as to perform $n = \frac{(k+1)B}{(k+1)C_1+C_2}$ steps with $m = \frac{B}{(k+1)C_1+C_2}$ number of participants.
Then variance:

$$\frac{\sigma^2}{\frac{B}{(k+1)C_1+C_2}} \frac{1+\rho^{k+1}}{1-\rho^{k+1}}$$

The goal is to minimize variance. Let's look on the next function of $k$:

$$f(k) = \frac{\sigma^2}{\frac{B}{kC_1+C_2}} \frac{1+\rho^{k}}{1-\rho^{k}}$$

.

It has minimum when $k$ is a solution for the following equation.

$$2C_1 \log(\rho)\rho^k k - C_1\rho^{2k} + 2C_2 \log(\rho)\rho^k + C_1 = 0$$

I don't know if there is explicit expression for the solution.
Check if the second derivative is always positive.
Now, let's imagine that I have graph and I know $\rho$. I try to find $k$ experimentally and using $k$ from equation.
I will take as $\rho$ covariance between neighbors on the graph with field. =( It did not work (only in ER graph, but I am not sure)
General case
Variance of mean in general case

$$var\left[\bar{X}\right] = \frac{1}{n}\sum_{i=2}^{r} \frac{1-\lambda_i^2 - 2\frac{\lambda_i}{n} + 2\frac{\lambda_i^{n+1}}{n}}{(1-\lambda_i)^2} < f, v_i >_\pi^2$$

Variance of mean simplified

$$var\left[\bar{X}\right] = \frac{1}{n}\sum_{i=2}^{r}\frac{1+\lambda_i}{1-\lambda_i} < f, v_i >_{\pi}^{2}$$

Variance of mean with skipping

$$var\left[\bar{X}^k\right] = \frac{1}{n}\sum_{i=2}^{r}\frac{1+\lambda_i^k}{1-\lambda_i^k} < f, v_i >_{\pi}^{2}$$

Function: variance of mean with skipping having fixed budget and payments (simplified)

$$f(k) = \frac{1}{\frac{B}{kC_1+C_2}}\sum_{i=2}^{r}\frac{1+\lambda_i^k}{1-\lambda_i^k} < g, v_i >_{\pi}^{2}$$

Function: variance of mean with skipping having fixed budget and payments (general case)

$$var\left[\bar{X}\right] = \frac{1}{n}\sum_{i=2}^{r}\frac{1-\lambda_i^2-2\frac{\lambda_i}{n}+2\frac{\lambda_i^{n+1}}{n}}{(1-\lambda_i)^2} < f, v_i >_{\pi}^{2}$$

$$\sigma_{\hat{\mu}}^2(k) = \frac{\sigma^2}{\frac{B}{kC_1+C_2}}\frac{1+\rho^k}{1-\rho^k}$$

.

$$\sigma_{\hat{\mu}}^2(k) = \frac{1}{\frac{B}{kC_1+C_2}}\sum_{i=2}^{r}\frac{1+\lambda_i^k}{1-\lambda_i^k} < g, v_i >_{\pi}^{2}$$

$$\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$$

$$\sigma_{\hat{\mu}}^2(k) = \frac{\sigma^2}{n}\frac{1+\rho^k}{1-\rho^k}$$