DOI: 10.5455/2320-6012.ijrms20131106

Research Article

Does smoking delay pregnancy? Data analysis by a tweaked geometric distribution answers

Ramalingam Shanmugam*

School of Health Administration, Texas State University, San Marcos, TX 78666, USA

Received: 14 July 2013
Accepted: 4 August 2013

***Correspondence:**
Dr. Ram Shanmugam,
E-mail: rs25@txstate.edu

ABSTRACT

Background: Smoking is generally known to be carcinogenic and health hazardous. What is not clear is whether the smoking impacts on the woman's reproductive process. There have been medical debates on whether a woman in the child bearing age may delay her pregnancy due to smoking. A definitive conclusion on this issue has not been reached perhaps due to a lack of appropriate data evidence. The missing link to answer the question might be exercising a suitable model to extract the pertinent data information on the number of missed menstrual cycles by *smoking* women versus *non-smoking* women. This article develops and demonstrates a statistical methodology to answer the question.

Methods: To construct such a needed methodology, a new statistical distribution is introduced as an underlying model for the data on the number of missed menstrual cycles by women who smoke. This new distribution is named *Tweaked Geometric Distribution* (TGD). Several useful properties of the TGD are derived and explained using a historical data in the literature.

Results: In the data of 100 smokers and 486 non-smokers, on the average, *smoking women* missed 3.22 menstrual cycles and non-smoking women missed only 1.96 menstrual cycles before becoming pregnant. The smoking women exhibited more variation than the non-smoking women and it suggests that the non-smoking women are more *homogeneous* while the smoking women are more *heterogeneous*. Furthermore, the *impairment level to pregnancy due to smoking* among the 486 women is estimated to be 5% in a possible scale of zero to one. The 5% impairment level appears like a small amount, but its impact can be felt once it is cast in terms of *fecundity*. What is fecundity? The terminology *fecundity* refers the chance for a woman to become pregnant. The fecundity is 0.24 for smoking woman while it is 0.34 for non-smoking woman. The fecundity of a non-smoking woman is more than twice the fecundity of a smoking woman.

Conclusion: The smoking is really disadvantageous to any one in general and particularly to a woman who wants to become pregnant.

Keywords: p-value, Statistical power, Prevalence, Hypothesis test, Nuisance parameter

INTRODUCTION

Let us begin with the concept called *menstrual cycle* in this medical research. What is it? It is a cycle of changes in uterus and ovary of a fertile female as a sign of preparation for reproduction. The average number of days among the human females is 28 per menstrual cycle. The cycle is controlled by the *endocrine system*. Any hormonal change could interfere with this control system to delay or even prevent the reproductive process. A question to ask is then: "Does smoking cause hormonal changes and consequently delay the pregnancy?" An answer is hidden in the pertinent data. The extraction of data information requires an appropriate model and methodology. The literature does not offer a suitable model and methodology. This need is fulfilled in this article.

Let us discuss the history of human smoking. The smoking can be traced to at least 5,000 B. C. if not to an earlier time.[1] The *World Health Organization* estimated that 5.4 million deaths in 2004 alone were caused by smoking. The smoking contains carcinogenic pyrolytic compounds which trigger genetic mutations. Several medical studies have proven beyond any reasonable doubt that smoking is a leading cause of many diseases like heart disease, lung cancer, erectile dysfunction, birth defects etc. among others. Many governments currently deter their citizens from smoking by a heavy taxation of the products, increased health insurance premium, and anti-smoking campaign. Still, fertile women in the child bearing age brackets smoke without realizing its harmful effects on their hormonal changes which cause a significant delay to become pregnant.

Before proceeding further, let us look at a historical data[2] in Table 1. It is clear from the data configurations of the smokers versus non-smokers in Figure 1, the smoking is a *delay factor* to become pregnant. With this clue, the next step is to discover an underlying model for the data. A follow up step is to construct a way to analyze the data to bring out the *impact of smoking* on the number of missed menstrual cycles, Y before the pregnancy occurs? To quantify such a latent impact of smoking on Y, we need an appropriate underlying model for the data and it is taken up first. The model is named *tweaked geometric distribution* (TGD) as the well-known geometric distribution[3] is a particular case of the TGD. The statistical properties, the estimators and hypothesis testing of the parameters of TGD are derived and utilized with the data in Table 1. A few conclusive thoughts are mentioned in the end of this article. Though our demonstration in this article focuses on human pregnancy due to smoking as a delay factor, the TGD is more versatile to explain similar scenarios with delay factors in economic, public health, engineering, social or business studies.

Table 1: # missed menstrual cycles to pregnancy among smokers and non-smokers.

Y = *m*	**# smokers**	**# non-smokers**	$Pr(Y\ ^{3}\ m \,\vert\, \theta, smokers)$	$Pr(Y\ ^{3}\ m \,\vert\, \theta, non\_smokers)$
0	29	198	1	1
1	16	107	0.763	0.66
2	17	55	0.582	0.44
3	4	38	0.444	0.29
4	3	18	0.339	0.19
5	9	22	0.259	0.13
6	4	7	0.197	0.08
7	5	9	0.151	0.06
8	1	5	0.115	0.04
9	1	3	0.088	0.02
10	1	6	0.067	0.02
11	3	6	0.051	0.01
12	7	12	0.039	0.01
Mean	3.22	1.965		
Variance	9.646	2.807		
$\hat{\theta}_{mle,\hat{\phi}} =$	0.28	0.34		
$\hat{\phi}_{mle} =$	0.05	0		
Fecundity	0.24	0.34		
P-value for	$H_0 : \phi = 0$ is 0.0009			
Power with	$H_1 : \phi^* = 0.1$ is 0.99			

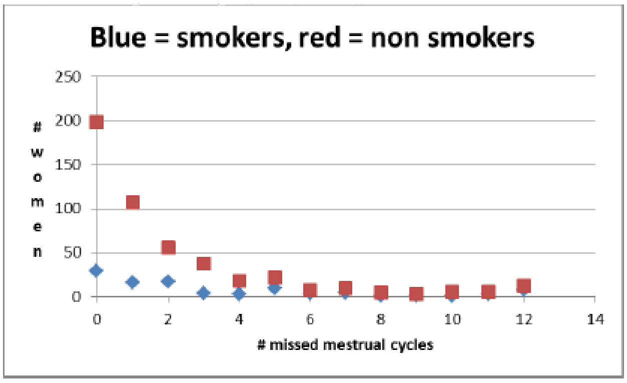

Figure 1: The pattern of pregnancy among smokers versus non-smokers.

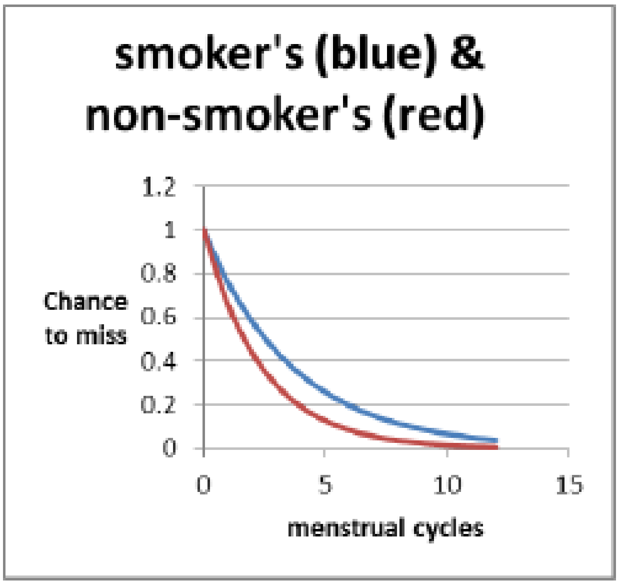

Figure 2: Survival function of smoker versus non-smoker.

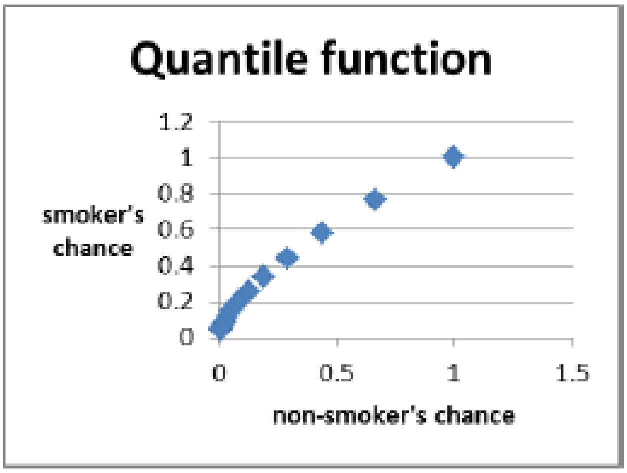

Figure 3: Survival function of smoker's in terms of survival function of non-smoker's.

METHODS

Suppose that a random variable, Y denote the number of *missed menstrual cycles* until the pregnancy occurs. The sample support for Y is the collection C of non-negative integers. The usual geometric distribution is a natural candidate to describe the probability pattern of Y if the probability for a fertile female to become pregnant in a menstrual cycle is an unknown parameter $0 < \theta < 1$. This means

$$\Pr(Y = y|\theta) = (1-\theta)^y \theta; y = 0,1,2,...,. \qquad (1)$$

This suggests that the expected number of missed menstrual cycles to pregnancy is

$$E(Y = y|\theta) = \sum_{y=0}^{\infty} yP(Y = y|\theta) = (\frac{1}{\theta} - 1) \qquad (2)$$

Rewriting (2) as $\theta = \frac{1}{1 + E(Y = y|\theta)}$, it helps to recognize that the probability, θ for a fertile female to become pregnant in a menstrual cycle decreases when the expected number of missing menstrual cycles increases. With this interpretation, let us look at the results in Table 1. The sample average number of missed menstrual cycles is 1.96 for non-smokers while it is 3.22 for smokers. The sample average estimates the expected number. The smoker's average is about 1.63 times more than the non-smoker's average. Also, the smoker's variance is about 3.43 times more than the non-smoker's variance. The variance is a measure of predictability. Lesser variance is indicative of higher predictability.

These remarks suggest that the probability, θ for a fertile female to become pregnant in a menstrual cycle *must* be impaired by the smoking. So, how do we construct an appropriate model?

The average number of missed menstrual cycles by the smoking women is almost twice the average by the non-smoking women. This data evidence is a clue about an effect of smoking on the endocrine system. Only an appropriate underlying model for the data could disclose it.

For this purpose, let $\phi \geq 0$ be an unknown *impairment* level due to smoking. That is to mention that the smoker's probability to become pregnant in a menstrual cycle ought to be $\frac{\theta - \phi}{1 - \phi}$ such that $\phi < \theta < 1 - \phi$. Assume that the probability pattern of the missing menstrual cycles with the impairment is still geometric. Hence, a counter-part of (1) for the smokers is then

$$\Pr(Y = y|\theta, \varphi) = (\frac{1-\theta}{1-\phi})^y (\frac{\theta - \phi}{1-\phi}); \qquad (3)$$
$$y = 0,1,2,...; 0 < \theta < 1; \phi < \theta < 1 - \phi.$$

The probability mass function in (3) is named *tweaked geometric distribution* (TGD). When a catalyst such as the smoking is absent, that is $\phi = 0$, the TGD reduces to the geometric distribution in (1). In this sense, the geometric distribution is nested within the TGD in (3).

The expected value of the TGD in (3) is

$$E(Y = y|\theta,\varphi) = (\frac{1-\theta}{\theta-\phi}). \qquad (4)$$

Notice that the expected value in (4) reduces to the expression (2) when $\phi = 0$. In other words, the TGD in (3) is more versatile than the geometric distribution in (1) for data like in the Table 1 or similar ones in other applications.

It is easily verified that the expected value in (4) is larger than its counterpart expected value in (2). This suggests that the expected value of the TBD with the presence of impairment due to smoking is larger than the expected value under the absence of the smoking effect. This explains the reason for the sample average number of missed menstrual cycles is 1.96 for the non-smokers while it is 3.22 for the smokers. Hence, the TGD in (3) is appropriate for the data in Table 1.

The variance of the TGD in (3) is

$$\text{var}(Y = y|\theta,\varphi) = \frac{(1-\phi)(1-\theta)}{(\theta-\phi)^2} \qquad (5)$$

which is greater than the variance $\text{var}(Y = y|\theta) = \frac{(1-\theta)}{\theta^2}$ of the geometric distribution in (1). The variance is a measure of predictability. The smaller variance implies a greater predictability. In this sense, the missed menstrual cycles of a smoker offers a lesser predictability compared to its counterpart of a non-smoker. Another way of interpreting it is that with respect to missing menstrual cycles, the smokers constitute a *heterogeneous* group. The non-smokers constitute a *homogeneous* group.

Another useful property of the TGD in (3) is its survival function. The survival function portrays how likely a smoker might miss at least a specified m menstrual cycles before pregnancy occurs? The survival function is

$$\Pr(Y \geq m|\theta,\varphi) = (\frac{1-\theta}{1-\phi})^m \qquad (6.a)$$

for a smoking woman. The survival function[3] is

$$\Pr(Y \geq m|\theta,\varphi = 0) = (1-\theta)^m \qquad (6.b)$$

for a non-smoking woman. Comparing the expressions in (6.a) and (6.b), we notice that the chance for a smoking woman to miss m or more menstrual cycles is consistently higher than that of non-smoking woman. The plot of (6.a) in terms of (6.b) is called the quantile function. More convexity in the quantile function's configuration suggests more significant difference between the smoker's and non-smoker's groups and such a graphical judgment is quick and effective. A picture does worth thousand words. Furthermore, the expressions in (6.a) and (5.b) become a basis to address a conditional situation for both smoking and non-smoking woman. Given that a smoker has missed at least m menstrual cycles, how likely for her to miss *additional q* menstrual cycles before becoming pregnant? In probability language, the statement is

$$\Pr(Y \geq m+q|Y \geq m) = \frac{\Pr(Y \geq m+q)}{\Pr(Y \geq m)}$$
$$= (\frac{1-\theta}{1-\phi})^q = \Pr(Y \geq q)$$

for a smoking woman which suggests the existence of a lack of memory in the smoking woman's endocrinal system. Likewise, a non-smoking woman's endocrinal system possesses a *lack of memory* as well because

$$\Pr(Y \geq m+q|Y \geq m, \phi = 0)$$
$$= (1-\theta)^q = \Pr(Y \geq q|\phi = 0)$$

according to the geometric distribution in (1). A practical implication of this memory less property is that the clock counting the missed menstrual cycles could recalibrated to zero at the end of every menstrual cycle whether the woman is smoker or otherwise.

Furthermore, the chance for a smoking woman to become pregnant in a current menstrual cycle is referred as *fecundity*. The word fecundity is defined[4] as a woman's ability to conceive and carry a fetus to a full term to deliver a live birth. With the TGD in (3) for a smoking woman, we notice her fecundity as

$$\Pr(Y = 0|smo\ker) = (\frac{\theta-\phi}{1-\phi}). \qquad (7)$$

The *fecundity* for a non-smoker is

$$\Pr(Y = 0|non\_smo\ker) = \theta \qquad (8)$$

due to using geometric distribution in (1) for non-smoker. Notice that a smoker's fecundity is smaller than a non-smoker's fecundity because the expression (7) is smaller than the expression (8). The smoking women are in a disadvantage when it comes to become pregnant.

Having noticed the importance of TGD, it is time to develop the estimators of its parameters. The maximum likelihood estimators (MLE) are preferable because of its *invariance* property[3]. That is, the MLE of a function of the parameters is simply the function of the MLE of the parameters. Consider a random sample $y_1, y_2, ..., y_n$ from TGD in (3). Then, the log-likelihood function is

$$\ln L = n\ln(\theta-\phi) - n(1+\bar{y})\ln(1-\phi)$$
$$+n\bar{y}\ln(1-\theta) \qquad (9)$$

Solving simultaneously the equations $\partial_\theta \ln L = 0$ and $\partial_\phi \ln L = 0$, the MLE of the parameters in (10) and (11) are obtained. That is,

$$\hat{\phi}_{mle} = \frac{|s^2 - \bar{y}^2|}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|} \qquad (10)$$

and

$$\hat{\theta}_{mle,\hat{\phi}} = \frac{\bar{y} + |s^2 - \bar{y}^2|}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|} \qquad (11)$$

where $\bar{y}$ and s^2 are sample mean and variance respectively. The MLE[3] of the parameter of the geometric distribution in (1) is

$$\hat{\theta}_{mle,\phi=0} = \frac{1}{(\bar{y}+1)} \qquad (12)$$

Notice that when $s^2 \to \bar{y}^2$, not only the MLE in (10) but also the MLE in (11) approach respectively to zero and the MLE in (12). After estimating the *impairment parameter* using (10), of interest to the medical community must be whether the estimated impairment is statistically significant or negligible? An answer to this question requires testing the null hypothesis $H_0 : \phi = 0$ against an alternative hypothesis $H_1 : \phi = \phi^* > 0$ .The most powerful test for this purpose is the Wald's likelihood ratio test[5] (WLRT). Using the likelihood function in (9), after algebraic simplifications, the log-WLRT to test the null hypothesis is

$$-\ln \Re_* = -\ln L(\phi = 0, \hat{\theta}_{mle,\phi=0}) + \ln L(\hat{\phi}_{mle} \neq 0, \hat{\theta}_{mle,\hat{\phi}})$$
$$= (n\bar{y}+1)(\frac{|s^2 - \bar{y}^2|}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|}) \qquad (13)$$

When the research hypothesis is true, the log-WLRT becomes

$$-\ln \Re_* = (n\bar{y}+1)[\frac{|s^2 - \bar{y}^2|}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|} - \phi^*] \qquad (14)$$
$$-n\ln[(1-\phi^*)(1+|s^2 - \bar{y}^2|) - \phi^* \bar{y}]$$

The expressions (13) and (14) follow individually a non-central chi-squared distribution with a non-centrality parameter $\delta_0 = \frac{\hat{\phi}_{MLE}}{\mathrm{var}(\hat{\phi}_{MLE})}$ and $\delta_* = \frac{(\hat{\phi}_{MLE} - \phi^*)}{\mathrm{var}(\hat{\phi}_{MLE})}$ respectively where $\mathrm{var}(\hat{\phi}_{MLE})$ is a diagonal element of the variance-covariance matrix

$$\Sigma = [\begin{matrix} \mathrm{var}(\hat{\phi}_{MLE}) & \mathrm{cov}(\hat{\phi}_{MLE}, \hat{\theta}_{MLE}) \\ \mathrm{cov}(\hat{\phi}_{MLE}, \hat{\theta}_{MLE}) & \mathrm{var}(\hat{\theta}_{MLE}) \end{matrix}] = I^{-1}$$

which is the inverse of the *Fisher's information matrix*

$$I = E[\begin{matrix} -\partial^2_{\hat{\theta}\hat{\theta}} \ln L & -\partial^2_{\hat{\theta}\hat{\phi}} \ln L \\ -\partial^2_{\hat{\theta}\hat{\phi}} \ln L & -\partial^2_{\hat{\phi}\hat{\phi}} \ln L \end{matrix}].$$

After algebraic simplifications, we note that

$$I = n[\begin{matrix} \frac{(1-\phi)}{(\theta-\phi)^2(1-\theta)} & -\frac{1}{(\theta-\phi)^2}1 \\ -\frac{1}{(\theta-\phi)^2} & \frac{(1-\theta)}{(\theta-\phi)^2(1-\phi)} \end{matrix}]$$

whose determinant is zero. The regular inverse is therefore not possible because of the matrix's singularity. But, its generalized inverse[6] I^- of the singular matrix I is possible in the sense $\mathrm{I}\mathrm{I}^-\mathrm{I} = \mathrm{I}$. For an example, if A^- is a generalized inverse matrix of the matrix A, then $\mathrm{A}\mathrm{A}^-\mathrm{A} = \mathrm{A}$ and $\mathrm{A}^-\mathrm{A}\mathrm{A}^- = \mathrm{A}^-$. Such a generalized inverse in our discussion is

$$\Sigma = I^- = \frac{1}{n}[\begin{matrix} \frac{(1-\theta)}{(\theta-\phi)^2(1-\phi)} & 0 \\ 0 & 0 \end{matrix}]$$

Hence, the estimate of the non-centrality parameter is

$$\hat{\delta}_* = \frac{n(\hat{\phi}_{mle} - \phi^*)(\hat{\theta}_{mle} - \hat{\phi}_{mle})^2(1-\hat{\theta}_{mle})}{(1-\hat{\phi}_{mle})}$$
$$= n(\frac{\bar{y}}{\bar{y}+1})[\frac{\bar{y}}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|}] \qquad (14)$$
$$[\frac{|s^2 - \bar{y}^2|}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|} - \phi^*]$$

Under the null hypothesis, the non-centrality parameter is $\hat{\delta}_0$ with $\phi^* = 0$ and $\hat{\delta}_*$ under the research hypothesis. However, it is well known[3] that a non-central chi-squared distribution with a non-centrality parameter δ is approximately $(1+\frac{\delta}{1+\delta})$ times the central chi-squared distribution with $(\frac{[1+\delta]^2}{1+2\delta})$ degrees of freedom (DF).

It then means that the null hypothesis $\mathrm{H}_0 : \phi = 0$ can rejected in favor of the research hypothesis $H_1 : \phi = \phi^* > 1$ when

$$(n\bar{y}+1)(\frac{|s^2 - \bar{y}^2|}{\bar{y}(\bar{y}+1) + |s^2 - \bar{y}^2|})$$

exceeds its critical value $(1+\frac{\hat{\delta}_0}{1+\hat{\delta}_0})\ \chi^2_{(\frac{[1+\hat{\delta}_0]^2}{1+2\hat{\delta}_0})DF,\alpha}$ at a chosen significance level, α. In other words, the p-value to reject the null in favor of the research hypothesis is

$$p - value \approx \Pr[\chi^2_{(\frac{[1+\hat{\delta}_0]^2}{1+2\hat{\delta}_0})DF,\alpha} > \frac{(n\bar{y}+1)(\frac{|s^2-\bar{y}^2|}{\bar{y}(\bar{y}+1)+|s^2-\bar{y}^2|})}{(1+\frac{\hat{\delta}_0}{1+\hat{\delta}_0})}]. \tag{16}$$

The *statistical power* is the probability of accepting a true specific research hypothesis in an event $\phi^* = \phi_1 \neq 0$. That is, for a specified significance level, α

$$power \approx \Pr[\chi^2_{(\frac{[1+\hat{\delta}_1]^2}{1+2\hat{\delta}_1})DF} < \frac{(1+\frac{\hat{\delta}_0}{1+\hat{\delta}_0})}{(1+\frac{\hat{\delta}_1}{1+\hat{\delta}_1})} \frac{(-\ln \Re_*)}{(-\ln \Re_0)} \chi^2_{(\frac{[1+\hat{\delta}_0]^2}{1+2\hat{\delta}_0})DF,\alpha}]. \tag{17}$$

RESULTS & DISCUSSION

In this section, we examine how the methodology of Section2 works with data in Table 1. There were 100 smokers and 486 non-smoking women who missed the menstrual cycles 0, 1, 2,…or 12. According to (4), the estimated average number of menstrual cycles is 3.22 for smokers and 2.63 for non-smokers. The estimated chance for smokers and for non-smokers to miss at least 0, 1, 2,… menstrual cycles using (6.a), (6.b), (10), (11) and (12) in the Figure 2. Notice that the smokers dominate consistently the non-smokers with delays.

The graph of (6.a) for smokers in terms of (6.b) for non-smokers is shown in Figure 3. The convexity in the configuration of graph in Figure 3 portrays graphically the significant difference between the smokers and non-smokers. A picture does worth the thousand words.

The fecundity, according to (7) and (8), is 24% for smokers and 34% for non-smokers with a difference of 10%. According to (10), the estimated impairment due to smoking is 5% which is statistically significant with the p-value = 0.0009 according to (16). This is to be interpreted that the null hypothesis $H_0 : \phi = 0$ is rejected in favor of the alternative hypothesis $H_1 : \phi = \phi^* > 0$ with 99.91% confidence level.

According to (17), the statistical power of accepting a specific *true* alternative hypothesis $H_1 : \phi^* = 0.1$ is 99% which is really good. This confirms that our methodology of Section 2 works well.

CONCLUSIONS

We witnessed that the smoking is hazardous to women who desire to become pregnant. Based on the methodology, the impairment to pregnancy due to smoking can be estimated and tested for its statistical significance. The usage of data with the methodology confirms that the impairment does indeed significantly delay the missing of menstrual cycles for the smoking women.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Gilman SL, Xun Z. Smoke: A global history of smoking, edited by Gilman SL, Xun Z. Reaktion Books Ltd., London, 2004.
2. Weinberg CR, Gladen BC. The beta geometric distribution applied to comparative fecundability studies. Biometrics 1986; 42: 547-560.
3. Stuart A, K. Ord. Kendall's Advanced Theory of Statistics. 1st Edn., Griffin Publication, London, U. K., 1994.
4. Last JM. A Dictionary of Epidemiology. Oxford University Press, Oxford, U. K., 2001.
5. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society 1943; 54: 426-482.
6. Schott JR. Matrix Analysis for Statistics, John Wiley Press, Hoboken, N. J., 2005.

DOI: 10.5455/2320-6012.ijrms20131106
Cite this article as: Shanmugam R. Does smoking delay pregnancy? Data analysis by a tweaked geometric distribution answers. Int J Res Med Sci 2013;1:343-8.