Tutorial: Binomial distribution in Stata

Using Stata to calculate binomial probabilities

Suppose X is a random variable that follows a binomial distribution; thus X represents the number of successes out of n trials with success probability p.

`binomialp(n,k,p)`	returns the probability of observing floor(k) successes in floor(n) trials when the probability of a success on one trial is p.
`binomial(n,k,p)`	returns the probability of observing floor(k) or fewer successes in floor(n) trials when the probability of a success on one trial is p.
`binomialtail(n,k,p)`	returns the probability of observing floor(k) or more successes in floor(n) trials when the probability of a success on one trial is p.

Example: Uzbeki Flour Fortification Program

In 2003, a flour fortification program was implemented in Uzbekistan to attempt to lower the rates of anemia among women of reproductive age. Before the program was implemented, the prevalence of anemia was 60%. In 2007, four years after implementing the fortification women, suppose 100 women of reproductive age were randomly selected to provide blood samples to test for anemia. Let X be the random variable denoting how many of the 100 women were anemic. Suppose that the prevalence of anemia in Uzbekistan did not change between 2003 and 2007.

1. Would the binomial distribution provide an appropriate model?

   B binary outcome
   I independent because women were randomly selected
   N sample size is fixed
   S same p

2. What is the expected number of women with anemia?

   $\mu = n * p = 60$

3. In a random sample of women in Uzbekistan, what is the typical departure of the number of women with anemia from this mean number?

$$
\begin{aligned}
sd(X) &= \sqrt{var(X)} \\
&= \sqrt{n * p * (1-p)} \\
&= \sqrt{100 * 0.6 * 0.4} \\
&= \sqrt{24} \\
&= 4.9
\end{aligned}
$$

4. What is the probability that exactly 60 women develop the disease? (use the formula)

$$\binom{n}{k} p^X (1-p)^{n-X} = \binom{100}{60} 0.6^{60} 0.4^{40} = 0.081$$

```
. di comb(100, 60)*0.6^60*0.4^40
.08121914
```

5. What is the probability that exactly 50 women are anemic?

```
. di binomialp(100, 50, 0.6)
.01033751
```

6. What is the probability that at least 50 women are anemic?

```
. di binomialtail(100, 50, 0.6)
.98323831
```

Alternatively, we could use the `binomial` command to calculate this probability, since $P(X > 50) = 1 - P(X \leq 49)$.

```
. di binomial(100, 49, 0.6)
.01676169
. di 1 - binomial(100, 49, 0.6)
.98323831
```

7. Now, assume that the prevalence of anemia actually dropped after implementation of the program, and the prevalence of anemia was 40% in 2007. Now, what is the probability that at least 50 women are anemic?

```
. di binomialtail(100, 50, 0.4)
.0270992
```

Note that under the assumption of no change in prevalence between 2003 and 2007, the probability that more than fifty women had anemia was very high. If the prevalence of anemia dropped to 40%, the probability that at least 50 women were anemia was then very low. So, if we collected data on 100 women and observed fewer than 50 cases of anemia, this would suggest that anemia prevalence dropped over time!