Geometric Distribution

Consider a random experiment that is closely related to the one used in the definition of a binomial distribution.

- ✓ **Again, assume a series of Bernoulli trials (*independent trials with constant probability p of success on each trial*).**
- ✓ **However, instead of a fixed number of trials, trials are conducted until a success is obtained.**

Let the random variable X denote the number of trials until the first success. In the transmission of bits, X might be the number of bits transmitted until an error occurs.

Example: The probability that a bit transmitted through a digital transmission channel is received in error is 0.1. Assume the transmissions are independent events, and let the random variable X denote the number of bits transmitted **until** the first error.

Then P(X=5) is the probability that the first four bits are transmitted **correctly** and the **fifth bit is in error**. This event can be denoted as

$\{OOOOE\}$, where O denotes an **okay** bit. Because the trials are independent and the probability of a correct transmission is 0.9,

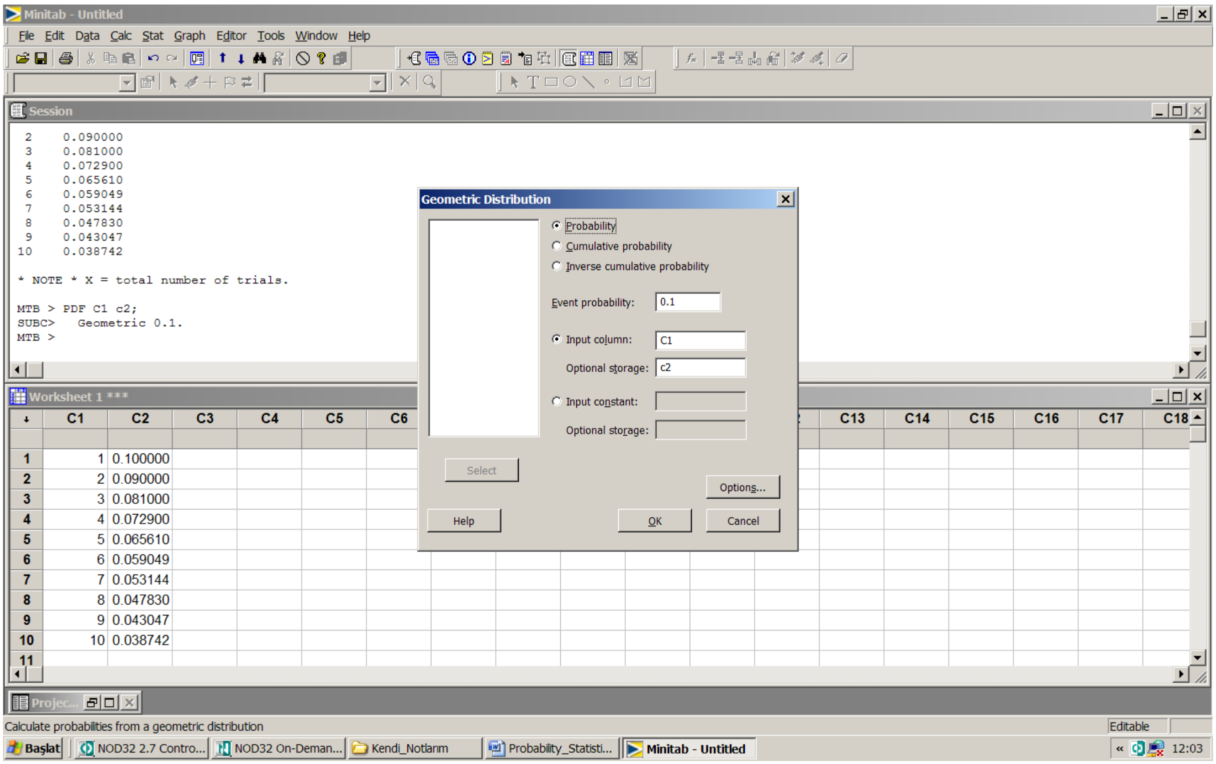
$$P(X=5)=P(OOOOE)=(0.9)^4(0.1)=0.066$$

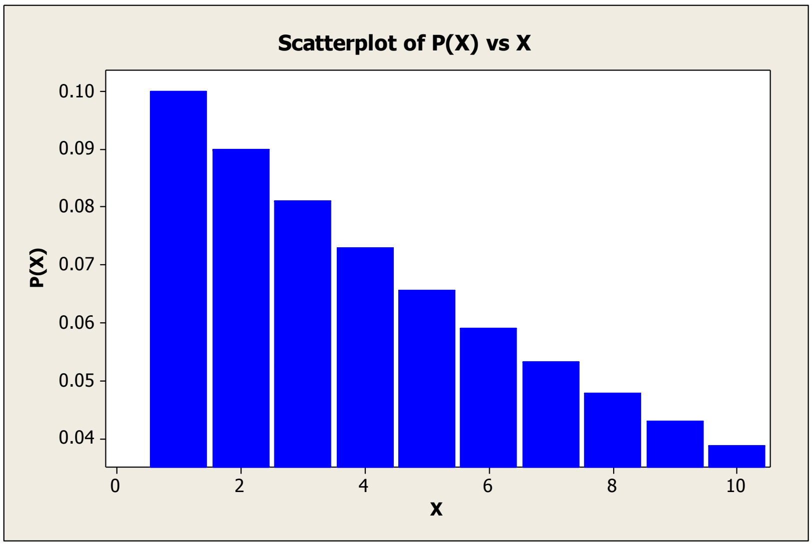
Note that there is some probability that X will equally any integer value. Also, if the first trial is a success, X=1. Therefore, the range of X is [1,2,3,….] that is all positive integers.

Definition:

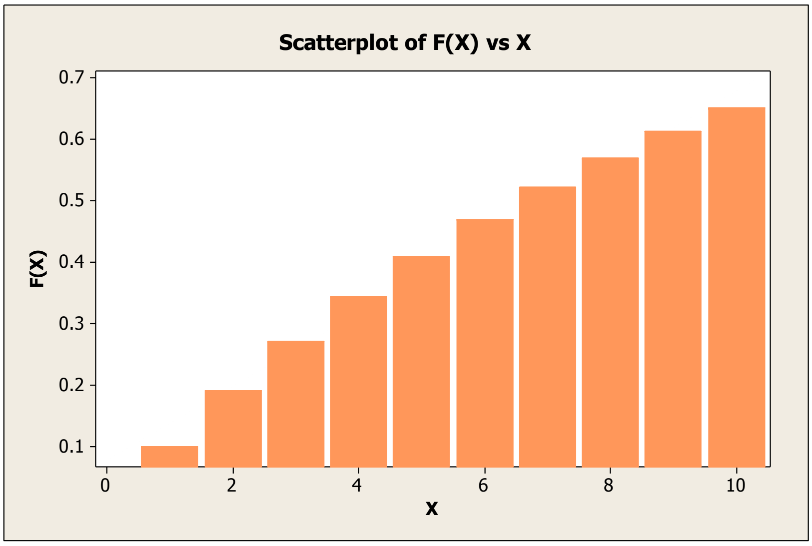
In a series of Bernoulli trials (independent trials with constant probability p of success), let the random variable X denote the number of trials until the first success. Then X is a geometric random variable with parameter 0 < p < 1 and

$$P(X = x) = (1 - p)^{x-1} p \qquad x = 1, 2, \ldots$$

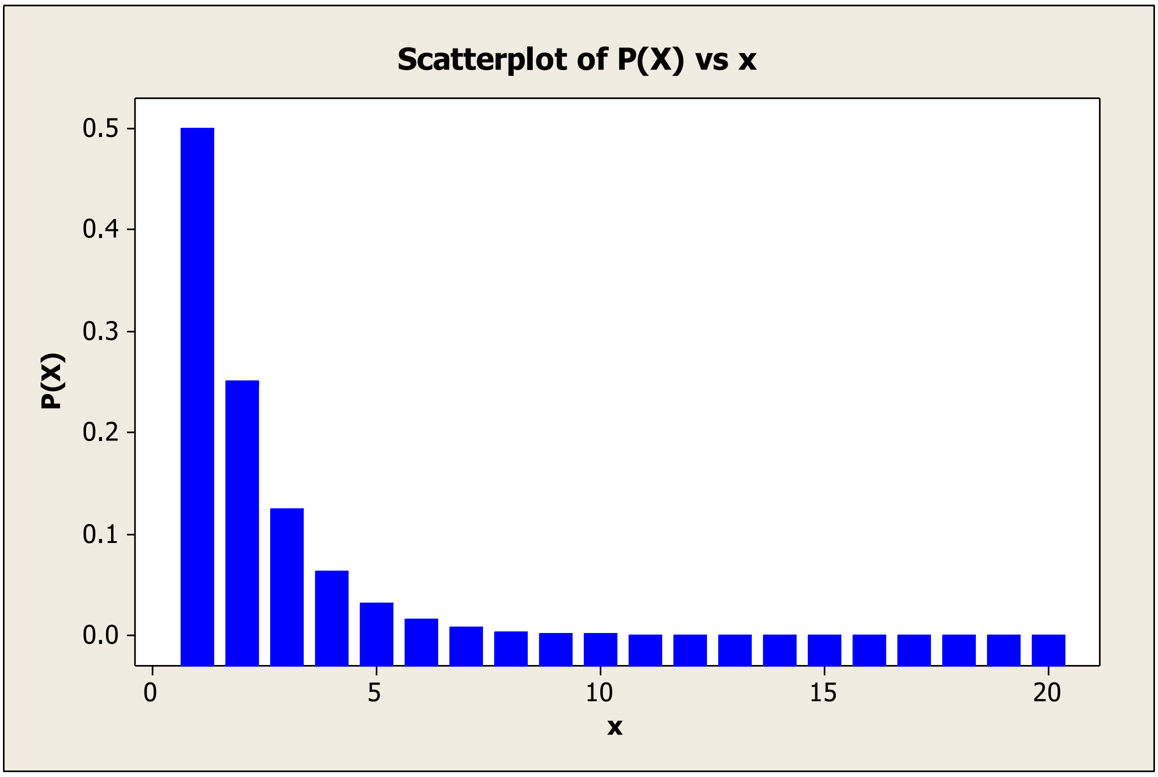




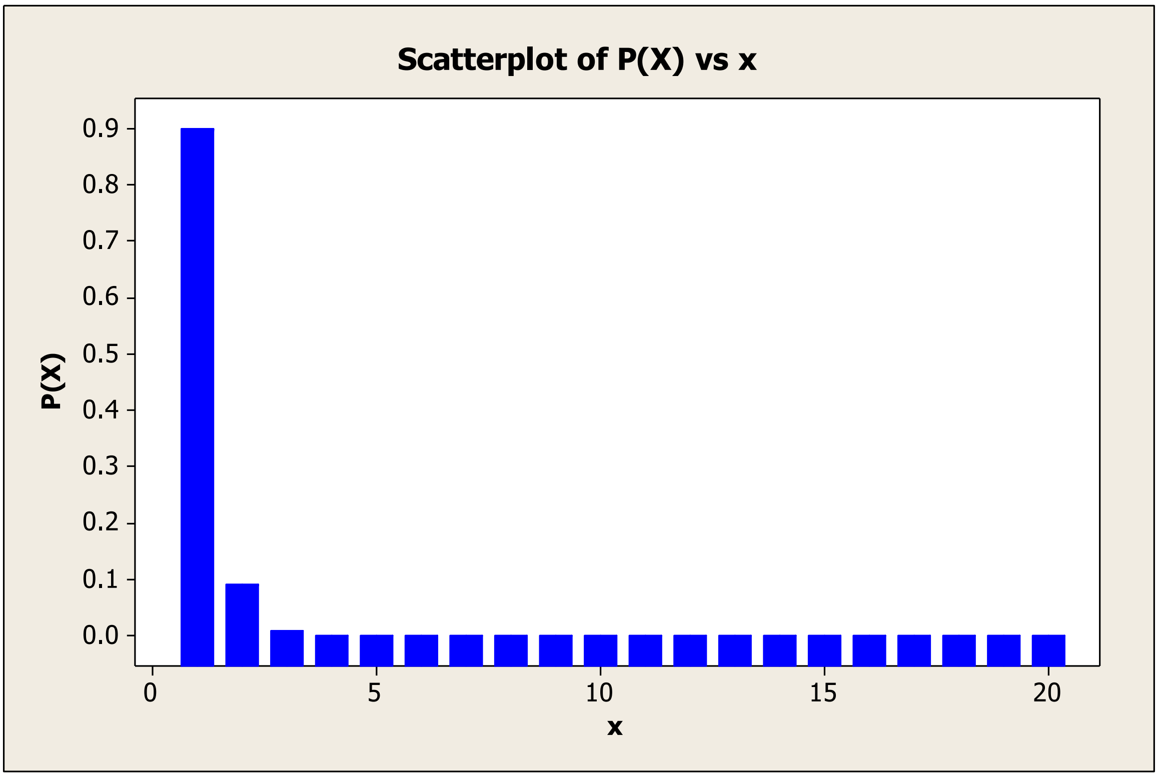
Geometric Probability Distribution with p=0.10 x=1…10



Geometric Cumulative Distribution with p=0.10 x=1…10



Geometric Probability Distribution with p=0.50 x=1…20



Geometric Probability Distribution with p=0.90 x=1…20

Mean and Variance of Geometric Distribution

If X is a geometric random variable with parameter p,

$$\mu = E(X) = 1/p \quad \text{and}$$

$$\sigma^2 = V(X) = (1-p)/p^2$$

$$\mu = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = p\sum_{x=1}^{\infty} xq^{x-1} \quad \text{where } q=1-p$$

The right hand side of the equation is recognized to be partial derivative with respect to q of

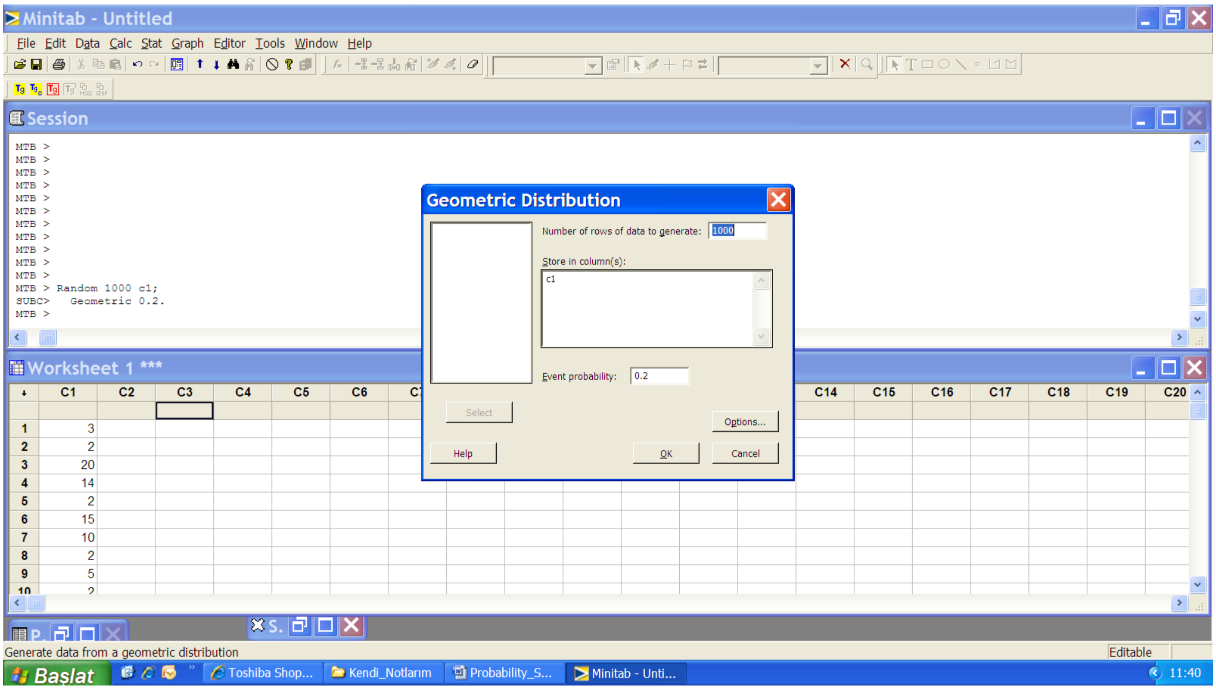
$$p\sum_{x=1}^{\infty} q^x = \frac{pq}{1-q}$$

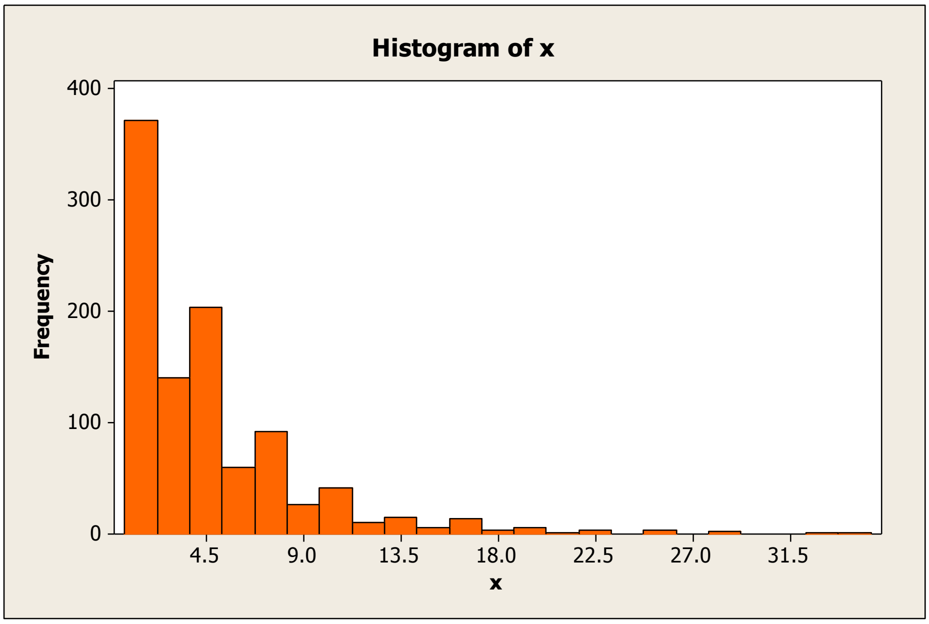
where the last equation is obtained from the known sum of geometric series. Therefore,

$$\mu = \frac{\partial}{\partial q}\left[\frac{pq}{1-q}\right] = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

and the mean is derived.

To obtain the variance, first derive $E(X^2)$ by a similar approach. This can be obtained from partial second derivatives with respect to q. Then $V(X)=E(X^2)-(E(X))^2$ is applied.






```
MTB > desc c1
```

Descriptive Statistics: x

```
Variable     N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
x         1000   0  4.675    0.137  4.320    1.000  2.000   3.000  6.000

Variable  Maximum
x          35.000
```

$$\mu = E(X) = 1/p = 1/0.2 = 5$$

$$\sigma^2 = V(X) = (1-p)/p^2 = (1-0.2)/0.2^2 = 20$$

Note: Generating geometric random variable

- *Generate independent Bernoulli(p) random variables $Y_1, Y_2, \ldots$; Let I be index of the first successful one so X=I*

or

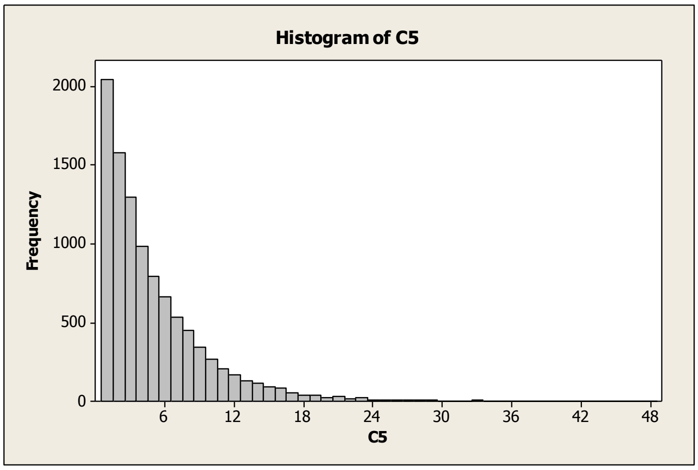
- *Generate U from U(0,1).*
- *Set* $X = 1 + \left\lfloor \frac{\ln(U)}{\ln(1-p)} \right\rfloor$. *(integer part)*

```
MTB > random 10000 c1;
SUBC> uniform 0 1.
MTB > let c2=loge(c1)/loge(0.8)
MTB > let c3=c2-0.5
MTB > round c3 c4
MTB > desc c4
Descriptive Statistics: C4

Variable        N  N*     Mean  SE Mean    StDev  Minimum       Q1
Median        Q3
C4          10000   0   3.9854   0.0445   4.4478   0.0000   1.0000
3.0000  6.0000
Variable  Maximum
C4        47.0000
MTB > let c5=c4+1
MTB > desc c5
Descriptive Statistics: C5

Variable        N  N*     Mean  SE Mean    StDev  Minimum       Q1
Median        Q3
C5          10000   0   4.9854   0.0445   4.4478   1.0000   2.0000
4.0000  7.0000

Variable  Maximum
C5        48.0000
```



Example: Consider the transmission of bits with the probability p=0.1. Find the mean and standard deviation of the geometric random variable.

The mean number of transmissions until the first error is

$$\mu = E(X) = 1/p = 1/0.1 = 10.$$

The standard deviation of the number of transmissions before the first error is

$$\sigma = [(1-p)/p^2]^{1/2} = [(1-0.1)/0.1^2] = 9.49.$$

Lack of Memory Property

A geometric random variable has been defined as the number of trials until the first success. However, because the trials are independent, the count of the number of trials until the next success can be started at any trial without changing the probability distribution of the random variable.

For example, in the transmission of bits, if **100 bits are transmitted,** the probability that the first error, after bit 100, occurs on bit 106 is the probability that the next six outcomes are **OOOOOE**. This probability is

$$\mathbf{(0.9)^5(0.1) = 0.059.}$$

This is identical to the probability that the initial error occurs on bit 6.

Negative Binomial Distribution

A generalization of a geometric distribution in which the random variable is the number of Bernoulli trials required to obtain <u>r successes results in the negative binomial distribution.</u>

Example: Digital Channel

Suppose that the probability that a bit transmitted through a digital transmission channel is received in error is 0.1. Assume the transmissions are independent events, and let the random variable X denote the number of bits transmitted until the **fourth error**.

Then, X has a negative binomial distribution with r=4 and p=0.1.

The P(X=10) is the probability that exactly three errors occur in the first nine trials is determined from the binomial distribution to be

$$\binom{9}{3}(0.1)^3(0.9)^6$$

Because the trials are independent, the probability that exactly three errors occur in the first 9 trials and trial 10 results in the fourth error is the product of probabilities of these, namely,

$$\binom{9}{3}(0.1)^3(0.9)^6.(0.1)=\binom{9}{3}(0.1)^4(0.9)^6=0.004464$$

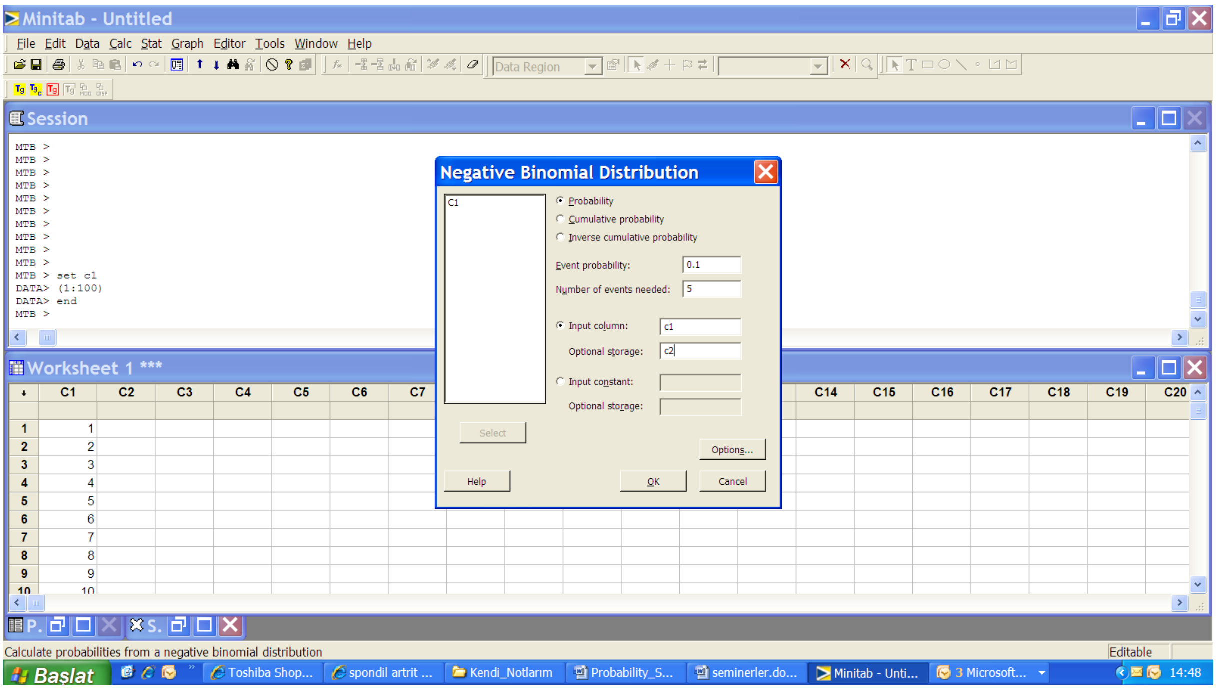
Definition:

In a series of Bernoulli trials (independent trials with constant probability p of a success), let the random variable X denote the number of trials until r successes occur. Then X is a negative binomial random variable with parameters 0<p<1 and r=1,2,3,4,…,and

$$P(X=x)=\binom{x-1}{r-1}(1-p)^{x-r}p^{r}$$

$$x=r,r+1,r+2,\ldots$$

- ***Because at least r trials are required to obtain r successes, the range of X is from r to ∞.***
- ***In the special case that r=1, a negative binomial random variable is a geometric distribution.***




```
MTB > pdf 10;
SUBC> negbinomial 0.1 4.
```

Probability Density Function

```
Negative binomial with p = 0.1 and
r = 4
 x    P( X = x )
10    0.0044641
```

```
* NOTE * X = total number of trials.
```

$$\binom{9}{3}(0.1)^3(0.9)^6.(0.1)=\binom{9}{3}(0.1)^4(0.9)^6=0.004464$$

```
MTB > set c1
DATA> (1:100)
DATA> end
MTB > PDF C1 c2;
SUBC>   NegBinomial 0.1 5.
MTB > PDF C1 c3;
SUBC>   NegBinomial 0.4 5.
MTB > print c1-c3
```

Data Display

Row	C1	C2	C3
1	1	0.0000000	0.000000
2	2	0.0000000	0.000000
3	3	0.0000000	0.000000
4	4	0.0000000	0.000000
5	5	0.0000100	0.010240
6	6	0.0000450	0.030720
7	7	0.0001215	0.055296
8	8	0.0002552	0.077414
9	9	0.0004593	0.092897
10	10	0.0007440	0.100329
11	11	0.0011160	0.100329
12	12	0.0015784	0.094596
13	13	0.0021308	0.085136
14	14	0.0027701	0.073785
15	15	0.0034903	0.061979
16	16	0.0042835	0.050710
17	17	0.0051402	0.040568
18	18	0.0060496	0.031830
19	19	0.0070003	0.024555
20	20	0.0079803	0.018662
21	21	0.0089779	0.013996
22	22	0.0099813	0.010374
23	23	0.0109794	0.007607
24	24	0.0119618	0.005525
25	25	0.0129187	0.003978
26	26	0.0138415	0.002842
27	27	0.0147223	0.002015
28	28	0.0155545	0.001419
29	29	0.0163322	0.000993
30	30	0.0170508	0.000691
31	31	0.0177066	0.000479
32	32	0.0182968	0.000330
33	33	0.0188196	0.000226
34	34	0.0192738	0.000154
35	35	0.0196593	0.000105
36	36	0.0199764	0.000071
37	37	0.0202261	0.000048
38	38	0.0204100	0.000032
39	39	0.0205300	0.000022
40	40	0.0205887	0.000014
41	41	0.0205887	0.000010
42	42	0.0205331	0.000006
43	43	0.0204250	0.000004
44	44	0.0202679	0.000003
45	45	0.0200652	0.000002
46	46	0.0198205	0.000001
47	47	0.0195373	0.000001
48	48	0.0192193	0.000001
49	49	0.0188699	0.000000
50	50	0.0184925	0.000000
51	51	0.0180905	0.000000
52	52	0.0176671	0.000000
53	53	0.0172254	0.000000

54	54	0.0167684	0.000000
55	55	0.0162989	0.000000
56	56	0.0158195	0.000000
57	57	0.0153327	0.000000
58	58	0.0148409	0.000000
59	59	0.0143462	0.000000
60	60	0.0138506	0.000000
61	61	0.0133560	0.000000
62	62	0.0128639	0.000000
63	63	0.0123760	0.000000
64	64	0.0118935	0.000000
65	65	0.0114178	0.000000
66	66	0.0109498	0.000000
67	67	0.0104906	0.000000
68	68	0.0100410	0.000000
69	69	0.0096017	0.000000
70	70	0.0091734	0.000000
71	71	0.0087564	0.000000
72	72	0.0083512	0.000000
73	73	0.0079582	0.000000
74	74	0.0075776	0.000000
75	75	0.0072096	0.000000
76	76	0.0068542	0.000000
77	77	0.0065115	0.000000
78	78	0.0061814	0.000000
79	79	0.0058640	0.000000
80	80	0.0055591	0.000000
81	81	0.0052665	0.000000
82	82	0.0049861	0.000000
83	83	0.0047176	0.000000
84	84	0.0044608	0.000000
85	85	0.0042155	0.000000
86	86	0.0039813	0.000000
87	87	0.0037579	0.000000
88	88	0.0035451	0.000000
89	89	0.0033426	0.000000
90	90	0.0031499	0.000000
91	91	0.0029667	0.000000
92	92	0.0027928	0.000000
93	93	0.0026278	0.000000
94	94	0.0024713	0.000000
95	95	0.0023230	0.000000
96	96	0.0021826	0.000000
97	97	0.0020498	0.000000
98	98	0.0019241	0.000000
99	99	0.0018054	0.000000
100	100	0.0016933	0.000000

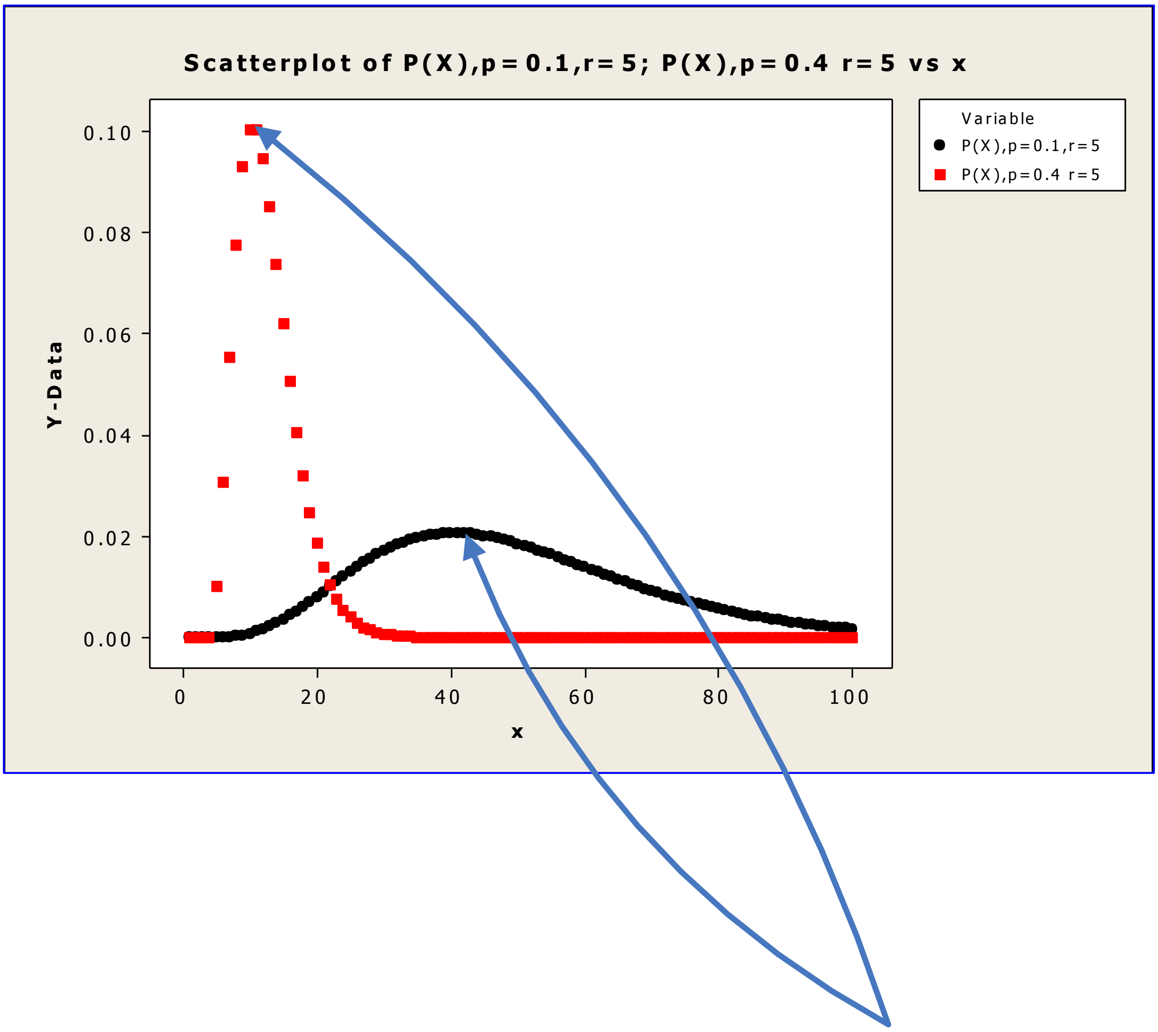
```
MTB > sum c2
```

Sum of C2

```
Sum of C2 = 0.976289
MTB > sum c3
```

Sum of C3

```
Sum of C3 = 1
```


Scatterplot of P(X),p=0.1,r=5; P(X),p=0.4 r=5 vs x
Y-Data
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
x
Variable
P(X),p=0.1,r=5
P(X),p=0.4 r=5

Compare

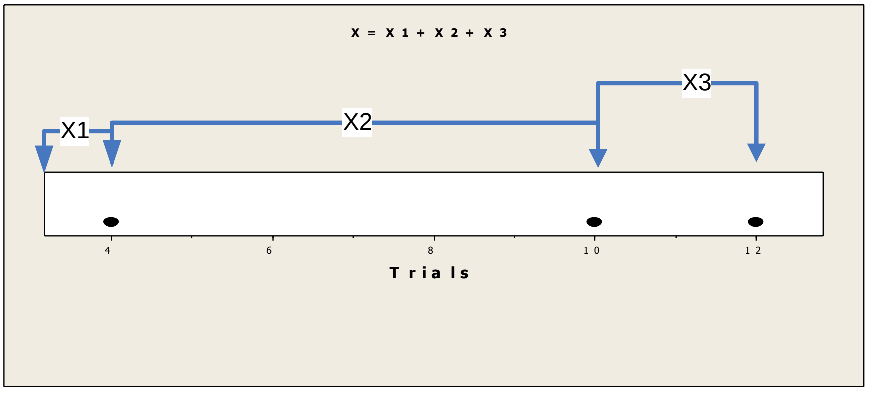
The lack of memory property of a geometric random variable implies the following.

- ***Let X_1 denote the number of trials required to obtain first success.***
- ***Let X_2 denote the number of trials required to obtain second success.***
- ***Let X_3 denote the number of trials required to obtain third success.***

Then, the total number of trials required to obtain r successes is

$$X=X_1+X_2+X_3+\ldots+X_r$$

- ***Because of the lack of memory property, each of the random variables, $X_1,X_2,\ldots,X_r$ has a <u>geometric distribution with the same value of p</u>.***
- ***<u>Consequently, a negative binomial random variable can be interpreted as the sum of r geometric random variables</u>.***



Mean and Variance of Negative Binomial Distribution

If X is a negative binomial random variable with parameter p and r,

$$\mu = E(X) = r/p$$ and

$$\sigma^2 = V(X) = r(1-p)/p^2$$

Example:

Let X_1 denote the number of trials required to obtain first success with p=0.1.

Let X_2 denote the number of trials required to obtain second success with p=0.1.

Let X_3 denote the number of trials required to obtain third success with p=0.1.

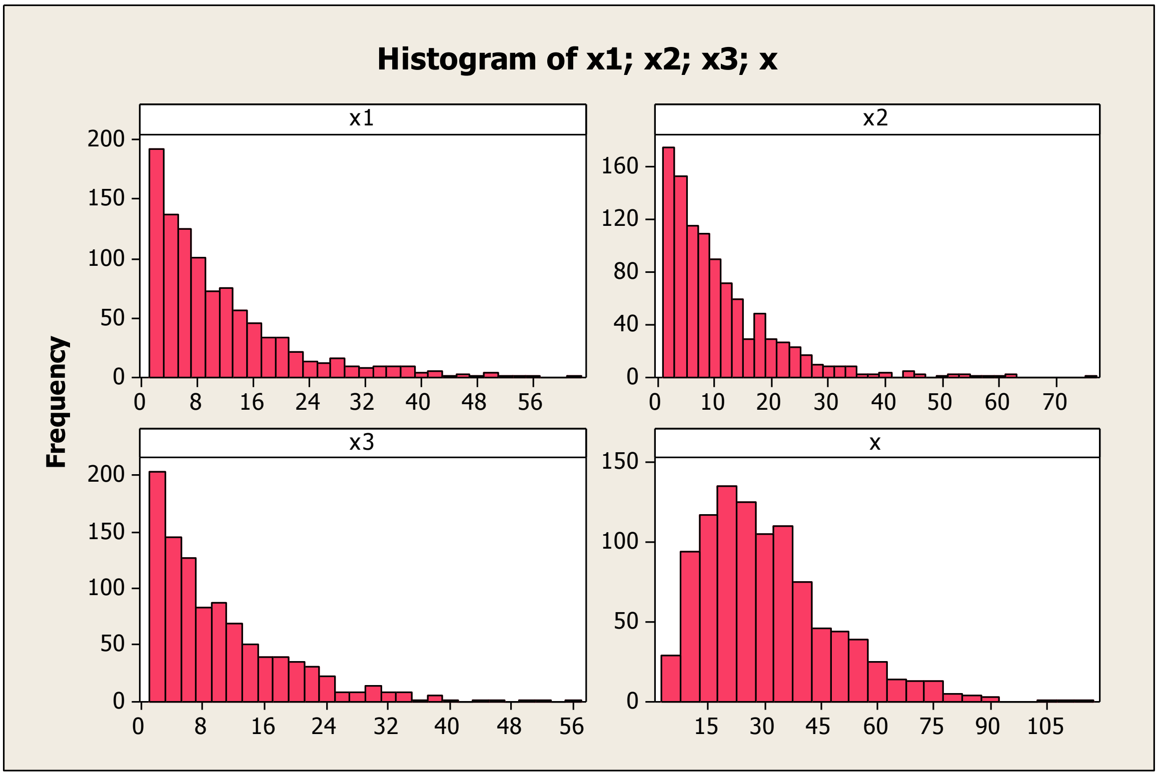
i	X_1	X_2	X_3	$X=X_1+X_2+X_3$
1	4	17	1	22
2	12	23	11	46
3	4	20	8	32
4	1	6	2	9
5	2	23	2	27
6	1	1	7	9
7	4	10	1	15
8	22	5	5	32
9	8	8	1	17
10	20	2	10	32
.	.	.	.	.
.	.	.	.	.
995	10	8	22	40
996	1	5	7	13
997	13	10	5	28
998	19	4	10	33
999	1	1	32	34
1000	1	28	5	34
Total	**10536**	**10416**	**9807**	**30759**

$$\hat{p}_1 = 1000/10536 = 0.0949$$

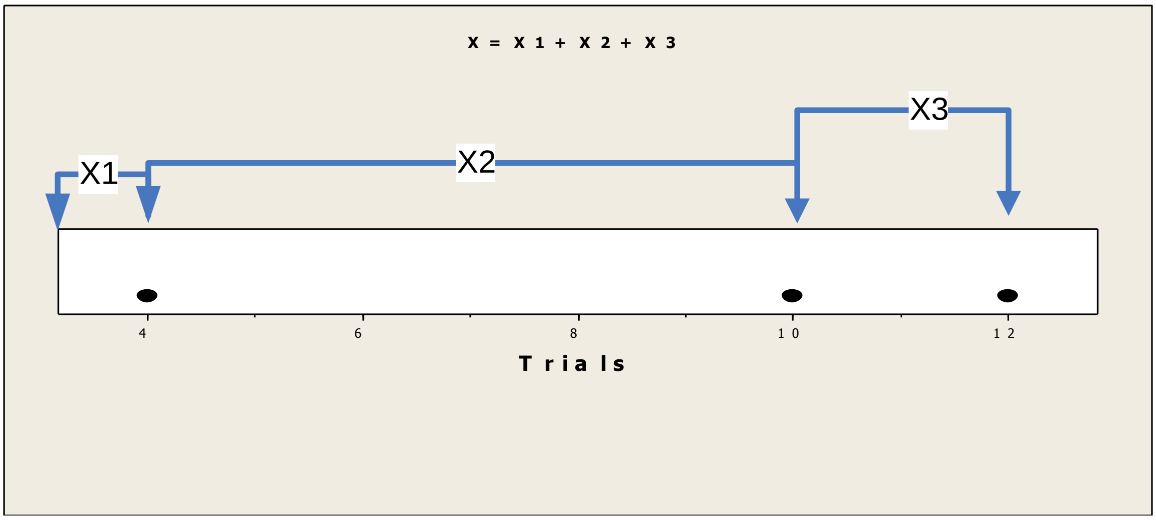
$$\hat{p}_2 = 1000/10416 = 0.0960$$

$$\hat{p}_3 = 1000/9807 = 0.10197$$

Generation of Table ...!



Let X_1=4 ,X_2=6,X_3=2 then X=4+6+2=12



Empirical results

Descriptive Statistics: x1; x2; x3; x

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
x1	1000	0	10.536	0.317	10.017	1.000	3.000	7.000	14.000
x2	1000	0	10.416	0.312	9.863	1.000	3.000	8.000	14.000
x3	1000	0	9.807	0.283	8.954	1.000	3.000	7.000	14.000
x	1000	0	30.759	0.549	17.363	3.000	18.000	27.000	40.000

Variable	Maximum
x1	61.000
x2	76.000
x3	56.000
x	114.000

MTB >

Theoretical results

$$\mu = E(X) = r/p = 3/0.1 = 30$$

$$\sigma^2 = V(X) = r(1-p)/p^2 = 270$$

$$\sigma = 16.43$$

Example: Web Servers

A web site contains three identical computer servers. Only one is used to operate the site, and the other two are spares that can be activated in case the primary system fails. The probability of a failure in the primary computer (or any activated spare system) from a request for service is 0.0005. Assuming that each request represents an independent trial, **what is the mean number of requests until failure of all three servers?**

- **Let X denote the number of requests until all three servers fail,**
- **Let X_1, X_2 and X_3 denote the number of requests before a failure of the first, second and third servers used, respectively.**
- **The requests are assumed to comprise independent trials with constant probability of failure p=0.0005.**
- **A spare server is not affected by the number of requests before it is activated.**

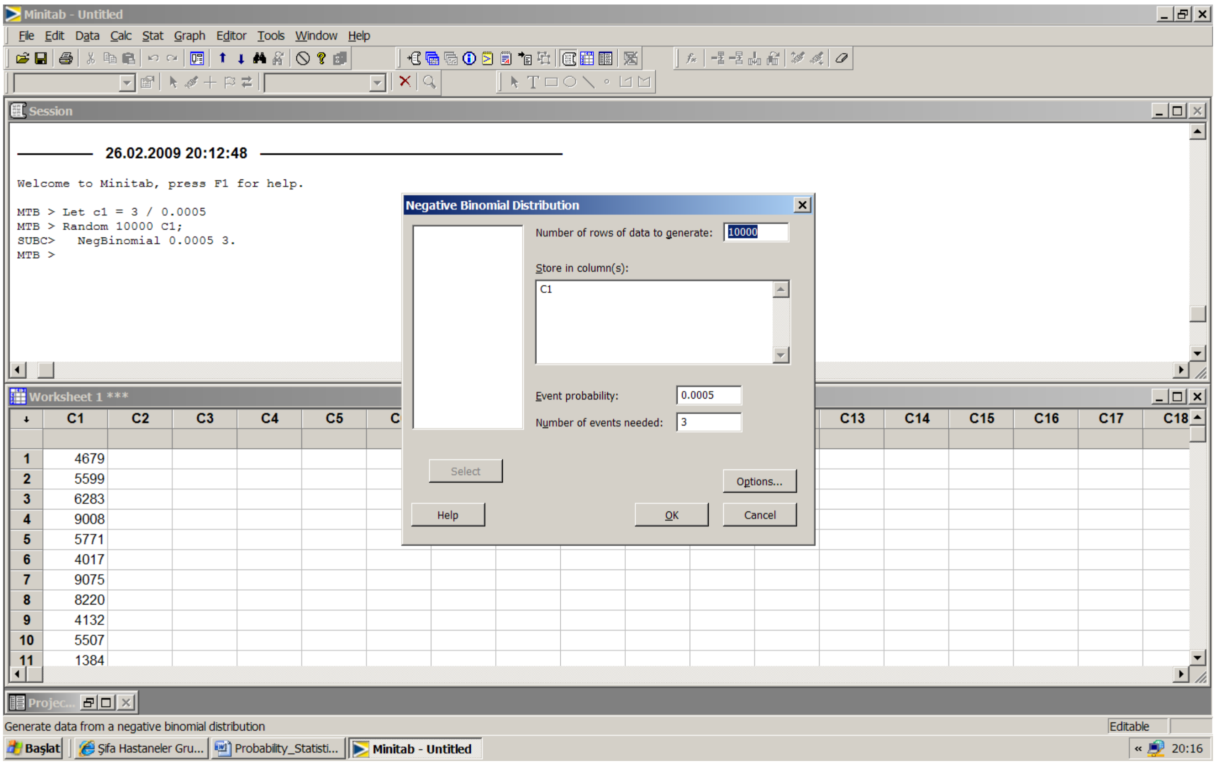
$X=X_1+X_2+X_3$

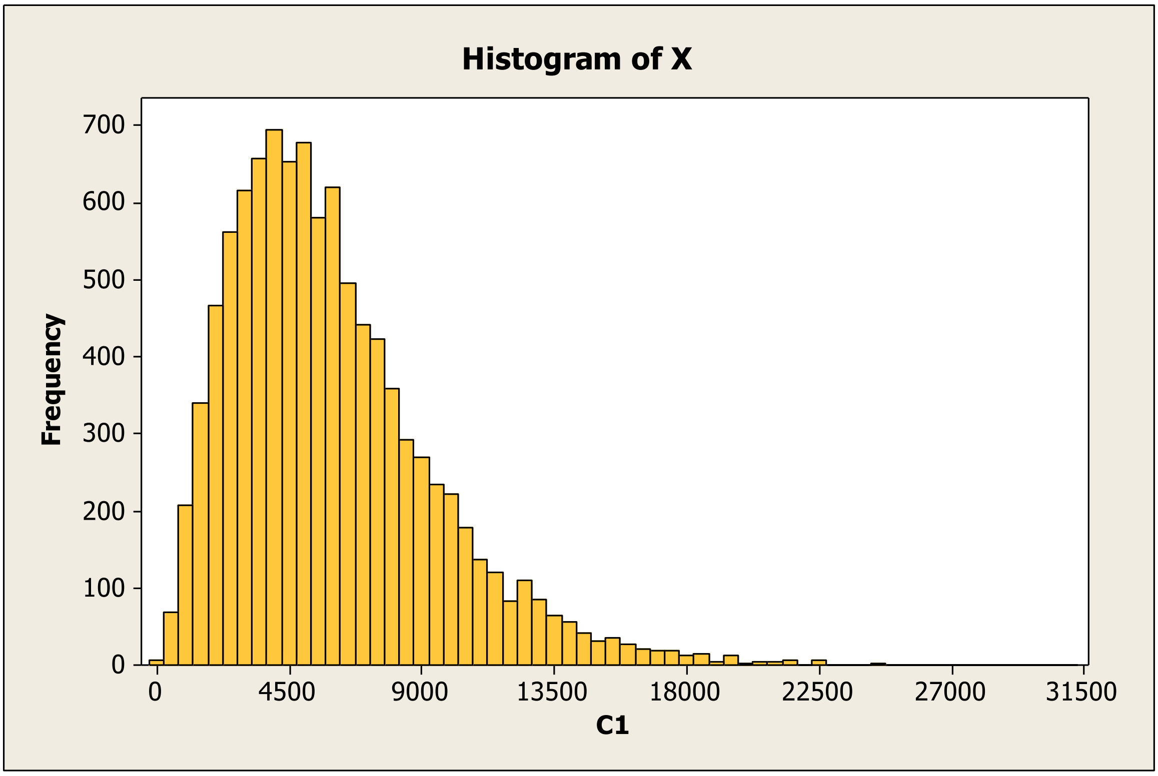
Therefore X has a negative binomial distribution with p=0.0005 and r=3.

$$P(X=x)=\binom{x-1}{r-1}(1-p)^{x-r}p^r$$

$$x=r,r+1,r+2,\ldots$$

Consequently E(X) =r/p = 3/0.0005 = 6000 requests.





Descriptive Statistics: x

```
Variable      N  N*    Mean  SE Mean   StDev  Minimum      Q1   Median      Q3
x         10000   0  5946.5     34.8  3475.1    107.0  3416.3   5287.5  7732.0

Variable  Maximum
x         31033.0
```

What is the probability that all three servers fail within five requests?

The probability is P (X≤5) and

P(X≤5)=P(X=3)+P(X=4)+P(X=5)

$$P(X=x)=\binom{x-1}{r-1}(1-p)^{x-r}p^{r}$$

$$x=r,r+1,r+2,\ldots$$

$$P(X\le 5)=(0.0005)^{3}+$$

$$\binom{3}{2}(0.0005)^{3}(0.9995)+$$

$$\binom{4}{2}(0.0005)^{3}(0.9995)^{2}$$

$$=1.25*10^{-10}+3.75*10^{-10}+7.49*10^{-10}=1.249*10^{-9}$$

```
MTB > CDF 5;
SUBC> NegBinomial 0.0005 3.
```

Cumulative Distribution Function

```
Negative binomial with p = 0.0005 and r = 3

x  P( X <= x )
5    0.0000000

* NOTE * X = total number of trials.
```

Inverse Cumulative Distribution Function

```
MTB > print c1 c2
```

Data Display

```
Row  C1    C2
  1   0  0.25
  2   1  0.50
  3   2  0.25
MTB > cdf c1 c3;
SUBC> disc c1 c2.
MTB > print c1-c3
```

Data Display

```
Row  C1    C2    C3
  1   0  0.25  0.25
  2   1  0.50  0.75
  3   2  0.25  1.00
```

```
MTB > invcdf 0.75;
SUBC> disc c1 c2.
```

Inverse Cumulative Distribution Function

```
Discrete distribution using values in C1
and probabilities in C2

x  P( X <= x )      x  P( X <= x )
0         0.25      1         0.75
```

```
MTB > pdf c1;
SUBC> negbinomial 0.2 2.
```

Probability Density Function

```
Negative binomial with p = 0.2 and r
= 2

x   P( X = x )
1        0.000
2        0.040
3        0.064

* NOTE * X = total number of trials.
```

```
MTB > cdf c1;
SUBC> negbinomial 0.2 2.
```

Cumulative Distribution Function

```
Negative binomial with p = 0.2 and r
= 2

x   P( X <= x )
1         0.000
2         0.040
3         0.104

* NOTE * X = total number of trials.
```

```
MTB > cdf 3;
SUBC> negbinomial 0.2 2.
```

Cumulative Distribution Function

```
Negative binomial with p = 0.2 and r
= 2

x  P( X <= x )
3        0.104

* NOTE * X = total number of trials.
```

$$P(X \leq ?) = 0.104$$

```
MTB > invcdf 0.104 ;
SUBC> negbinomial 0.2 2.
```

Inverse Cumulative Distribution Function

```
Negative binomial with p = 0.2 and r
= 2

x  P( X <= x )       x  P( X <= x )
3        0.104       4       0.1808

* NOTE * X = total number of trials.
```