Resampling Techniques and their Application

-Class 8-

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

Count Data

- Count data X_{ik} play an important role in biomedical sciences
- Endpoints are often realized as counts
- For example
  - Number of cells
  - Number of hospitalizations
  - Number of lesions
  - Number of patients
  - etc
- Why do we need extra methods?
- Response is always positive
- **Outcome is observed within a time frame**

Example

- Pediatric MS with disease onset under the age of 16 is uncommon and a rare disease
- Safety of interferon beta-1a compared to no treatment

Pediatric MS trial ($n_1 = n_2 = 8$)				
Endpoint	Group	Rate $\widehat{\lambda}_i$	Variance	Var/Mean
Lesions	Control	11.875	13.268	1.117
	Active	10.625	16.839	1.585
Relapses	Control	4.5	6.571	1.460
	Active	2.375	0.268	0.113

- Which distribution should be assumed?

```
#Lesions

x<-c(8,11,13,7,13,10,15,18)
y <-c(8,11,7,17,8,8,17,9)

#Relapses

x<-c(6,4,7,1,8,5,4,1)
y<-c(2,3,2,2,3,3,2,2)
```

Poisson Distribution

- The Poisson distribution is often a reasonable probability model for count data
- Count events that occur in time, space, volume,
  - number of typing errors per page made by a typist
  - number of radioactive particles that decay in a certain time period
  - number of bacterial colonies on a plate
- A random variable X has a Poisson distribution if the probability that X = k events will occur is given by

$$P(X = k) = \frac{\exp(-\lambda)\lambda^k}{k!}, k = 0, 1, \ldots$$

- λ is the average number of events per unit of time or area.
- Each choice of λ gives a different Poisson model.

Examples

- Ex 1: Mr. Dirty lives in a dorm room and lately he has noticed some cockroaches having fun in his room. He buys a roach-trap and puts it in his room. Let X = $\#$ of cockroaches captured in the roach-trap in a given time period. Suppose that X follows Poisson (λ = 3) distribution.
  - What is the average number of cockroaches that we expect in the roach-trap?
  - What is the probability that the trap will contain exactly 2 cockroaches?

$$P(X = 2) = \exp(-3)3^2/2! = 22.4\%$$

  - What is the probability the trap will contain at most 2 cockroaches?

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \exp(-3)(1 + 3 + 9/2) = 42.3\%$$

- Ex 2: Suppose the number of accidents at an intersection in a month follows Poisson(2.6) distribution. What is the probability that there are more than 3 accidents in a month?

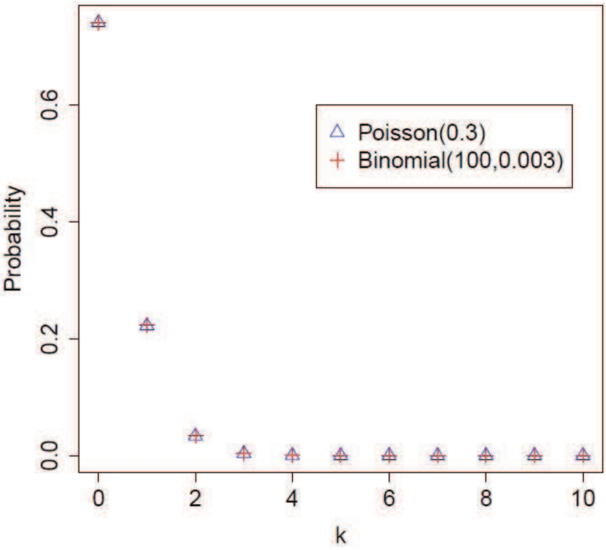
$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \exp(-2.6)(1 + 2.6 + 2.6^2/2 + 2.6^3/6) = 26.39\%$$

Poisson Approximation to Binomial Distribution

- Binomial gives probability of k successes in n trials while Poisson gives probability of k (rare) events occurring (with no upper bound on k). Is there a connection between these two?
- Think of "occurrence of event" as "success". And there is no failure!
- When p is small $p \leq 0.05$ and n is large ($n \geq 20$), the number of successes are rare and then Poisson (λ = np) $\approx$ Binomial(n, p).
- Example: Suppose the probability of mutation in the genome of a certain type of bacteria is 0.003 per generation. What is the probability that there will be at least 1 mutation in 100 generations for this type of bacteria?

$$\lambda = np = 100 \cdot 0.003 = 0.3 \Rightarrow P(X \geq 1) = 1 - P(X = 0) = 1 - \exp(-0.3) = 25.9\%$$

Poisson Approximation to Binomial Distribution



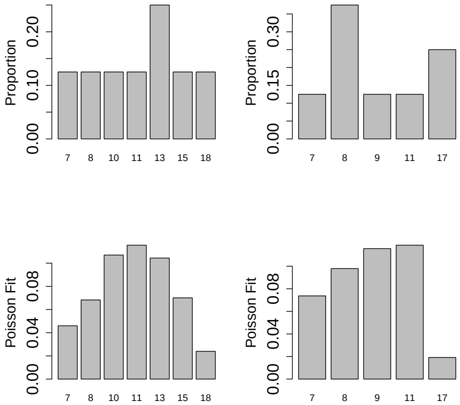
Characteristics of the Poisson Distribution

- Let $X \sim \mathit{Pois}(\lambda)$
  - $E(X) = \lambda$
  - $\mathit{Var}(X) = \lambda$
  - $E(X - E(X))^3 = \lambda$
  - What is the median of X?
- **Variance and mean are identical**

Example: Model Fit

- Model fit with T2 lesion data
- Check assumptions of poisson distribution
-
```
x<-c(8,11,13,7,13,10,15,18)
y <-c(8,11,7,17,8,8,17,9)
lx <- mean(x)
ly <- mean(y)

#Poisson fit
exp(-lx)*lx^x/factorial(x)
exp(-ly)*ly^y/factorial(y)
```



Negative Binomial Distribution

- Negative Binomial (NB) is another discrete distribution.
- The NB distribution is often a reasonable probability model for counts
- Number of trials in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of successes r occurs

Negative Binomial Distribution

- n: number of trials
- A random variable X has a NB distribution if the probability that X = k failures are needed to observe r successes is given by

$$P(X = k) = \left(\begin{array}{c} k + r - 1 \\ k \end{array} \right) p^r (1 - p)^k, k = 0, 1, \ldots$$

- Moments
  - $E(X) = r\frac{1-p}{p}$
  - $Var(X) = r\frac{1-p}{p^2}$
  - The variance is larger than the mean $\Rightarrow$ **overdispersion**
  - In the literature, term $\left(\begin{array}{c} k + r - 1 \\ k \end{array} \right) = \frac{\Gamma(r+k)}{k!\Gamma(r)}$. ($\Gamma(x)$: gamma function)

Negative Binomial Distribution

- **Parameterization**
- The above parameterization is a bit unhandy
- Easier to work with a rate λ as expectation as in the poisson model
- Define

$$
\begin{aligned}
E(X) &= \lambda = r\frac{1-p}{p} \Rightarrow p = \frac{r}{r+\lambda} \\
Var(X) &= r\frac{1-p}{p^2} = \lambda + \frac{\lambda^2}{r}
\end{aligned}
$$

```
#see Details in help file in R

?rnbinom


set.seed(1)
n=1000000
x=rnbinom(n,mu=10,size=3)
mean(x)
var(x)
10+10^2/3
```

Inference Methods

- Statistical model
  - X_{ik} count data with intensity rates λ_i and Variances σ_i^2, $i = 1, 2; k = 1, \ldots, n_i$
  - Count data are typically skewed, therefore log-transformation is applied
  - Hypothesis

$$H_0 : \log\left(\frac{\lambda_1}{\lambda_2}\right) = 0 \quad vs. \quad H_1 : \log\left(\frac{\lambda_1}{\lambda_2}\right) \neq 0$$

- We will use Poisson and Negative Binomial Regression
- Both models are used for count data
- Generalized Linear Models
- Functions *glm* and *glm.nb* in *R* (library MASS)

Example Evaluations

```
#Lesions

x<-c(8,11,13,7,13,10,15,18)
y <-c(8,11,7,17,8,8,17,9)
nx=length(x)
ny=length(y)

les=data.frame(res=c(x,y),
grp=factor(c(rep(1,nx),rep(2,ny))))
```

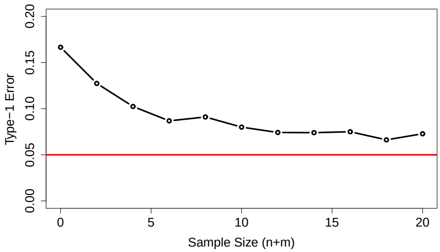
```
#Poisson
fitPois <- glm(res ~ grp, family="poisson", data=les)
summary(fitPois)
estPois <- cbind(Estimate = coef(fitPois), confint(fitPois))
SEPois=coef(summary(fitPois))[,2][2]

#Negative Binomial
library(MASS)
fit <- glm.nb(res ~grp , data = les)
summary(fit)
SENB=coef(summary(fit))[,2][2]
est <- cbind(Estimate = coef(fit), confint(fit))

log(mean(x))
log(mean(y)/mean(x))
```

Are the Methods Reliable?

- Simulation of NB Regression at $\alpha = 5\%$ level (10,000 simulation runs)
  - Generate data $X_{ik} \sim NB(10, 3)$, $k = 1, \ldots, n_i$; $i = 1, 2$
  - $n_i = 7 + m$, $m \in \{0, \ldots, 20\}$




```
simuNB<-function(m){
erg=c();n1=7+m;n2=7+m
grp=factor(c(rep(1,n1),rep(2,n2)))
for(i in 1:10000){
x=rnbinom(n1,mu=10,size=1/3)
y=rnbinom(n2,mu=10,size=1/3)
data=data.frame(res=c(x,y),grp=grp)
fit<-coef(summary(glm.nb(res~grp,data = data)))[,4][2]
erg[i] =(fit<0.05)}
result=data.frame(m=m,NBR=mean(erg))
result}
m=seq(0,20,1)
NB=matrix(0,ncol=2,nrow=length(m))
for(s in 1:length(m)){
NB[s,] = as.matrix(simuNB(m[s]))[1,]}
```

Are the Methods Reliable?

- The method is very liberal when sample sizes are small
- Applicability is doubtful
- Why?
  - Method uses Maximum Likelihood Estimators, which might be unstable when samples are small
  - The distribution of the test statistic does not involve variability of the variance estimators
  - Can we develop a new/improved version of the test?

Point Estimators

- Point estimators

$$\widehat{\lambda}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} X_{ik};\ \widehat{\sigma}_i^2 \quad = \quad \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(X_{ik} - \widehat{\lambda}_i)^2$$

- Variance of $log(\frac{\widehat{\lambda}_2}{\widehat{\lambda}_1})$ (obtained with Delta-Method*)

$$\tau^2 = Var\left(log\left(\frac{\widehat{\lambda}_2}{\widehat{\lambda}_1}\right)\right) = \frac{Var(\widehat{\lambda}_1)}{\lambda_1^2} + \frac{Var(\widehat{\lambda}_2)}{\lambda_2^2}$$

- Estimator

$$\widehat{\tau}^2 \quad = \quad \frac{\widehat{\sigma}_1^2/n_1}{\widehat{\lambda}_1^2} + \frac{\widehat{\sigma}_2^2/n_2}{\widehat{\lambda}_2^2}$$

- *Delta-Method: $Var(g(\widehat{\theta})) = [g'(\theta)]^2 Var(\widehat{\theta})$ for any estimator $\widehat{\theta}$ of θ and differentiable function g

Test Statistic

- Test statistic for $H_0 : log(\lambda_2/\lambda_1) = 0$

$$T = \frac{log(\widehat{\lambda}_2/\widehat{\lambda}_1)}{\widehat{\tau}}$$

- For large sample sizes, T follows $N(0, 1)$ distribution
- Explore resampling methods to approximate the test
- For example, permutation test

A Permutation Test

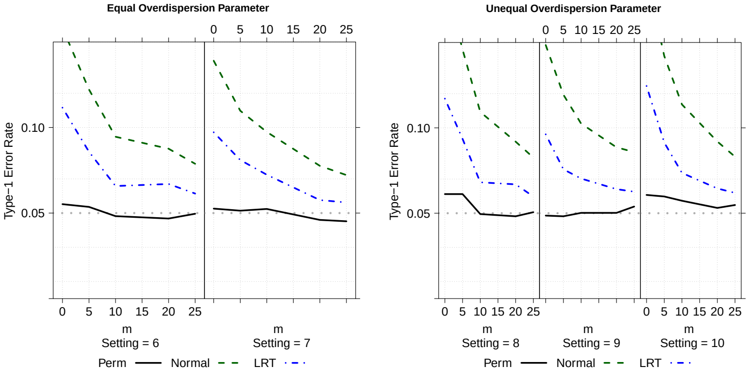
- Collect data in $\mathbf{X} = (X_{11}, \ldots, X_{2n_2})'$
  1. Randomly permute the data and obtain $\mathbf{X}^* = (X^*_{11}, \ldots, X^*_{2n_2})$
  2. Reassign $X^*_{11}, \ldots, X^*_{1n_1}$: group 1
  3. Remaining: group 2
  4. Compute $\widehat{\lambda}^*_1$, $\widehat{\lambda}^*_2$, $\widehat{\sigma}_1^{2,*}$, $\widehat{\sigma}_2^{2,*}$, $\widehat{\tau}^{2,*} = \frac{\widehat{\sigma}_1^{2,*}/n_1}{\widehat{\lambda}_1^{2,*}} + \frac{\widehat{\sigma}_2^{2,*}/n_2}{\widehat{\lambda}_2^{2,*}}$
  5. Compute $T^* = \frac{\log(\widehat{\lambda}^*_2/\widehat{\lambda}^*_1)}{\widehat{\tau}^*}$ and safe the value in T^*_ℓ
  6. Repeat the above a large number of times $(n_{perm} = 10K)$ and obtain $T^*_1, \ldots, T^*_{n_{perm}}$
  7. Compute the p-value as

$$p - Value = 2 * \min\left\{\frac{1}{n_{perm}} \sum_{\ell=1}^{n_{perm}} \mathcal{I}(T^*_\ell \leq T); \frac{1}{n_{perm}} \sum_{\ell=1}^{n_{perm}} \mathcal{I}(T^*_\ell \geq T)\right\}$$

- Shown: The permutation test is valid (Konietschke, F., Friede, T., Pauly, M. (2019))

Motivation: Count Regression

- Simulation
  - 2 groups, Negative Binomial data ($\lambda_1 = \lambda_2 = 10$); $n_i = 7 + m$, $m \in \{0, \ldots, 20\}$
  - Compare permutation test with NB binomial test (LRT) Note: Results obtained with individual time frames.



Example: Evaluation

Method	Effect	SE	Statistic	p-Value	95% CI
			T2 lesions		
Normal	0.111	0.174	0.638	0.524	(-0.231 ; 0.453)
Perm	0.111	0.174	0.638	0.545	(-0.269 ; 0.510)
NB-Reg	0.111	0.161	0.691	0.489	(-0.204 ; 0.428)
Pois-Reg	0.111	0.149	0.745	0.456	(-0.181 ; 0.405)
			Relapses		
Normal	0.639	0.216	2.964	0.003	(0.216 ; 1.062)
Perm	0.639	0.216	2.964	0.026	(0.116 ; 1.162)
NB-Reg	0.639	0.284	2.254	0.024	(0.096 ; 1.215)
Pois-Reg	0.639	0.284	2.254	0.024	(0.096 ; 1.215)

Appendix: Numerical Computation of the MLE

- For completeness, numerical solutions to compute the MLE are provided
- Using the parametrization from above, we have

$$P(X_{ik} = x_{ik}) = \frac{\Gamma(x_{ij} + r)}{x_{ij}!\Gamma(r)} \left(\frac{1/r\lambda_i}{1 + 1/r\lambda_i} \right)^{x_{ij}} \left(\frac{1}{1 + 1/r\lambda_i} \right)^{r}$$

- So, the probability to observe the data set is

$$\mathcal{L}(\beta_0, \beta_1, r) = \prod_{i=1}^{2} \prod_{k=1}^{n_i} \frac{\Gamma(x_{ij} + r)}{x_{ij}!\Gamma(r)} \left(\frac{1/r\lambda_i}{1 + 1/r\lambda_i} \right)^{x_{ij}} \left(\frac{1}{1 + 1/r\lambda_i} \right)^{r}$$

- In the GLM, we assume

$$E(X_{ik}|r, g_i) = \lambda_i = \exp(\beta_0 + \beta_1 g_i), \quad g_1 = 0; g_2 = 1$$

Appendix: Numerical Computation of the MLE

- Plugging-in the assumption, we get

$$\mathcal{L}(\beta_0, \beta_1, r) = \prod_{i=1}^{2}\prod_{k=1}^{n_i} \frac{\Gamma(x_{ij}+r)}{x_{ij}!\Gamma(r)} \left(\frac{1/r\exp(\beta_0+\beta_1 g_i)}{1+1/r\exp(\beta_0+\beta_1 g_i)}\right)^{x_{ij}} \left(\frac{1}{1+1/r\exp(\beta_0+\beta_1 g_i)}\right)^{r}$$

- So, we get for the log-likelihood

$$\begin{aligned}\ell(\beta_0, \beta_1, r) &= \sum_{i=1}^{2}\sum_{k=1}^{n_i} \log\left(\frac{\Gamma(x_{ij}+r)}{x_{ij}!\Gamma(r)}\right) + x_{ij}\log\left(\left(\frac{1/r\exp(\beta_0+\beta_1 g_i)}{1+1/r\exp(\beta_0+\beta_1 g_i)}\right)\right) \\ &- r\log\left(1+1/r\exp(\beta_0+\beta_1 g_i)\right)\end{aligned}$$

- Now, find the maximum using appropriate software, e.g. the *optim* function in *R*

Software Code

```
library(MASS)
x<-c(8,11,13,7,13,10,15,18)
y <-c(8,11,7,17,8,8,17,9)
xy<-c(x,y)
nx=length(x)
ny=length(y)

indx=c(rep(0,nx))
indy=c(rep(1,ny))
indxy=c(indx,indy)

maxi=function(parameter){
beta0=parameter[1]
beta1=parameter[2]
a2=parameter[3]
if(a2<=0){result=Inf}
else{
lambda= exp(beta0+beta1*indxy)
result=-1*(
sum(xy*log(1/a2*lambda/(1/a2*lambda+1))
-a2*log(1+1/a2*lambda)
+log(gamma(xy+a2))-log(factorial(xy)*gamma(1/a2^(-1)))))}}
parameters=optim(c(1,1,1),maxi)
parameters

my=mean(y)
mx=mean(x)
log(my/mx)
log(mean(x))
#close to the result using glm.nb
```

Appendix

Additional Material

Counting Techniques-I

- One basic process we commonly encounter is one that has two possible (mutually exclusive) outcomes.
  - E.g., head/tail, pass/fail, infected/not infected, positive/negative, defective/non-defective.
- Generic name: success/failure.
- Many chance-based questions for this type of process could be of interest such as:
  - In four tosses of a coin, what is the chance of getting exactly two heads?
  - In a lab experiment with a virus, what is the probability that half of the cells are infected?
  - If a blood test is done to 100 diseased individuals, what is the probability that 90 of them test positive?

Counting Techniques

- Enumeration becomes very cumbersome pretty soon (think 10 tosses of a coin!).
- We will learn in how many ways we can list k objects out of a total of n objects.
- There are two counting techniques:
    - **Permutation** = arrangement of objects into a sequence. Order of objects is important.
    - **Combination** = selection of objects. Order of objects is not important.
- Ex: We want to form a two-digit number using the digits 1, 2, 3, 4, 5 (without repeating)
- Permutation or combination? Why?
- Ex: We want to make a team of 3 players from a pool of 5 players. Permutation or combination? Why?

Permutation

- Ex: We want to form a two-digit number using the digits 1, 2, 3, 4, 5. Possible ways:

$$\begin{gathered}(1,2),(1,3),(1,4),(1,5)\\(2,1),(2,3),(2,4),(2,5)\\(3,1),(3,2),(3,4),(3,5)\\(4,1),(4,2),(4,3),(4,5)\\(5,1),(5,2),(5,3),(5,4)\end{gathered}$$

- In general, total $\#$ of permutations of k objects chosen (without replacement) from n distinct objects are

$$n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$$

Permutation-II

- Factorial notation: n! = n factorial = product of first n natural numbers = 1(2)(3)...(n - 1)(n)
- Define 0! = 1.
- Using factorial notation, our answer is $5 * 4 = 5!/3! = 20$
- Total $\#$ of permutations of k objects chosen from n distinct objects can be written compactly as

$$\frac{n!}{(n-k)!}$$

Combination

- Ex: How many two-player teams can be formed from a pool of 5 players?

$$(1,2),(1,3),(1,4),(1,5)$$
$$(2,3),(2,4),(2,5)$$
$$(3,4),(3,5)$$
$$(4,5)$$

- In general, total $\#$ of combinations of k objects chosen from n distinct objects =

$$\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$$

- The total $\#$ of combinations of k objects chosen from n distinct objects can be written compactly as

$$\frac{n!}{k!(n-k)!}$$

The Binomial Distribution

- Setting: Success (S) and Failure (F)
- Count the number of successes out of n trials
- Example: Toss a coin 4 times. What is the probability to observe $k = 2$ heads?
- We need two things:
  - How many possible ways we can get 2 heads in 4 tosses? The answer is

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

  - How likely is each of those possible ways?
- Consider one possible way: H H T T. For a coin with P(H) = 3/4, what is the probability of this event? Let's look at another possible way and its probability: $P(HHTT) = (\frac{3}{4})^2(\frac{1}{4})^2$

The Binomial Distribution

- We see each possible way has the same probability as the ordering does not matter:

- In general, if p denotes the P(Success), each possible way has probability of
  $p^2(1-p)^{n-k}$

- Thus combining answers to 1 and 2, we get the probability of getting exactly 2 heads in 4 tosses of coin as
  $6(\frac{3}{4})^2(\frac{1}{4})^2$

The Binomial Probability Distribution

- Binomial Distribution: In n independent "success-failure" trials with probability of success in each trial being p, the probability of getting k successes out of n trials is

$$\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}, k = 0, 1, \ldots, n$$

- Note that for each value of k (possible $\#$ of success), there is a probability. These probabilities add up to 1.
- Different values of n and p give different Binomial distributions. The assumptions of the Binomial distribution are
  - The trials must be independent.
  - The value of n must be fixed in advance.
  - p must be the same from trial to trial.

The Binomial Probability Distribution

- Ex: Roll a dice 3 times. Define success as getting 6. The probability distribution of $\#$ of sixes in 3 rolls is given by

$$\begin{aligned} k &= 0, 1, 2, 3 \\ P(X = k) &= \binom{3}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{3-k} \end{aligned}$$

- Thus, the probability of getting exactly 3 sixes is $\left(\frac{1}{6}\right)^3$ and getting at least 2 sixes is

$$P(X \geq 2) = P(X = 2) + P(X = 3) = 3 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right) + \left(\frac{1}{6}\right)^3 = 7.407\%$$

More Examples - II

- Ex: In a certain population, 25% of the adults experience hypertension at some point of their lives. Suppose 5 adults are randomly chosen from this population. What is the probability that 3 of them would have experienced hypertension?

$$P(X = 3) = \binom{5}{3} \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^2 = 8.8\%$$

- Ex: Consider tasting Baco Tell chalupas at a Baco Tell location. Suppose that 80% of the chalupas at this location taste good. You order 10 chalupas from this Baco Tell location. What is the probability that at least nine of them taste good?

$$P(X \geq 9) = P(X = 9) + P(X = 10) = 10 \left(\frac{8}{10}\right)^9 \left(\frac{2}{10}\right) + \left(\frac{8}{10}\right)^{10} = 37.58\%$$

Geometric Distribution

- Setting: Succes/Failure (same as Binomial distribution).
  - Repeat a sequence of independent success-failure trials until you get the first success. E.g., Roll a dice until you get a six
  - Geometric distribution provides the probability that the first success occurs at mth trial. If p is the probability of success in a trial, this probability is given by

$$P(X = m) = (1 - p)^{m-1} \cdot p$$

- Geometric vs. Binomial
  - For Binomial distribution the total number of trials is fixed while for Geometric distribution it is not fixed.
  - In other words, in Geometric probability model, there is no upper limit on m unlike in Binomial model.

Example

- Example: In a certain population, 25% of the adults experience hypertension at some point of their lives. Suppose we randomly select adults from that population. What is the probability that the fifth person is the first who has hypertension?

$$P(X = 5) = \left(\frac{3}{4}\right)^4 \left(\frac{1}{4}\right) = 7.9\%$$

- In the above example, what is the probability that we have to observe at least 5 people to find the first person with hypertension?

$$P(X \geq 5) = \{\text{First 4 are F}\} = \left(\frac{3}{4}\right)^4 \quad (= 1 - P(X \leq 4))$$

- In general, the probability that at least m trials are needed to get the first success is

$$P(X \geq m) = \{\text{First (m-1) are F}\} = (1 - p)^{m-1}$$

Example - Contd

- Suppose first three people are observed and there was no one with hypertension among them. Now, what is the probability that we have to observe at least five more people to observe the first person with hypertension? This is the conditional probability: P(at least 5 more trials needed before observing the first success $\mid$ first 3 trials resulted in failures).

$$P(X \geq 8|\text{First 3 are F}) = \frac{P(X \geq 8, \text{First 3 are F})}{P(\text{First 3 are F})} = \frac{P(X \geq 8)}{P(\text{First 3 are F})} = \frac{(1-p)^7}{(1-p)^3} = \left(\frac{3}{4}\right)^4$$

- How does it compare with the P(at least 5 trials needed to observe the first success) calculated earlier?

Lack of Memory Property

- How about P(at least 5 more trials needed to observe the first success $|$ first 10 trials resulted in failures)?

$$P(X \geq 15 | \text{First 10 are F}) = \frac{\left(\frac{3}{4}\right)^{14}}{\left(\frac{3}{4}\right)^{10}} = \left(\frac{3}{4}\right)^{4}$$

- This means the process resets itself! It has "no memory" of what happened before, and it always starts afresh after any sequence of consecutive failures.

- Given that the first k trials have all resulted in failures, the probability that at least m more trials are needed to get the first success is $(1 - p)^{m-1}$, which is same as the probability that at least m trials are needed to get the first success, when counted from the beginning.

Lack of Memory Property - Contd

- Given that the first k trials have all resulted in failures, the probability that the first success occurs at (k +m)th trial is $(1 - p)^{m-1}p$, which is same as the probability that the first success occurs at mth trial, when counted from the beginning.

Examples

A cereal manufacturer puts a special prize in 1/20 of the boxes.

- What is the probability that a prize is in the third box?

$$P(X = 3) = \left(\frac{19}{20}\right)^2 \left(\frac{1}{20}\right)$$

- What is the probability that you have to purchase at least 3 boxes to get a prize?

$$P(X \geq 3) = \left(\frac{19}{20}\right)^2$$

- Suppose you have already purchased 5 boxes and didn't get a prize. What is the probability that you have to purchase at least 3 more boxes to get a prize?

$$P(X \geq 8 | \text{First 5 are F}) = \frac{(19/20)^7}{(19/20)^5} = \left(\frac{19}{20}\right)^2$$

- What is the probability of getting a prize before the 3rd box?

$$P(X \leq 2) = P(X = 1) + P(X = 2) = 1/20 + 19/20 \cdot 1/20 = 9.75\%.$$

Negative Binomial Distribution

Example: Take a standard deck of cards, shuffle them, and choose a card. Replace the card and repeat until you have drawn two aces. Y is the number of draws needed to draw two aces. As the number of trials is not fixed (i.e. you stop when you draw the second ace), this makes it a negative binomial distribution.

Example: A representative from the National Football League's Marketing Division randomly selects people on a random street in Kansas City, Kansas until he finds a person who attended the last home football game. Let p, the probability that he succeeds in finding such a person, equal 0.20. Now, let X denote the number of people he selects until he finds $r = 3$ who attended the last home football game. What is the probability that X = 10?

Negative Binomial Example

- An oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil. What is the probability that the first strike comes on the third well drilled?

- What is the probability that the third strike comes on the seventh well drilled?

- What is the mean and variance of the number of wells that must be drilled if the oil company wants to set up three producing wells?

Characteristics

- **Binomial**
- $E(X) = np$
- $Var(X) = np(1-p)$

- **Poisson**
- $E(X) = \lambda$
- $Var(X) = \lambda$

- **Negative Binomial**
- $E(X) = r(1-p)/p \equiv \lambda$
- $Var(X) = r(1-p)/p^2 \equiv \lambda + \lambda^2/r$