Autumn 1392

Statistical analysis of data

Mehrnush Forutan

Department of Animal Science
Isfahan University of Technology

Slide 2

Data analysis

Steps that can be distinguished when analyzing data:

a) Exploratory analyses

1) Initial examination of the data.
2) Relations explanatory variables and the response variable.
3) Relations among explanatory variables.
4) Conclusions based on the exploratory analysis.

b) Building the model

c) Model criticism

d) Analysis using the final model.

Slide 3

a.1) Initial examination of the data - Why??

1. Asses the structure of the data

   What is the sample size?
   How many variables are there?
   What type of variables (continuous, categorical) are there?

2. Data quality

   Are there missing observations?
   How were missing values treated?
   Are there any outliers?

Slide 4

Response variable (Y)

Continuous, binomial, count data…
Is a general linear model appropriate?

Explanatory variables (X)

- all quantitative - linear regression models
- all qualitative - analysis of variance models
- quantitative and qualitative - analysis of covariance

fixed effects
co-variables

Slide 5

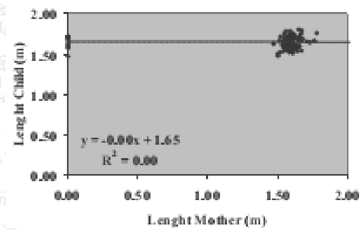
Missing data - check if there is a pattern.

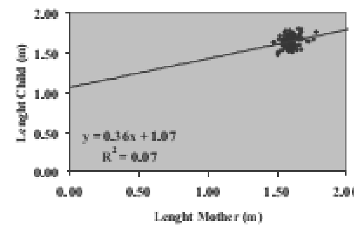
- If the missingness is completely unrelated to the response or the predictor variables, there is no problem.
- If there is a pattern in the missing data, however, things become complicated.

Slide 6

Missing data - how are they coded?

e.g. a missing value code of 0 or -99 for body weight of a cow will not be noticed by the software package but might have a big impact on the results.





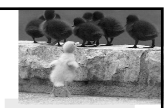
Slide 7

Exploratory analysis

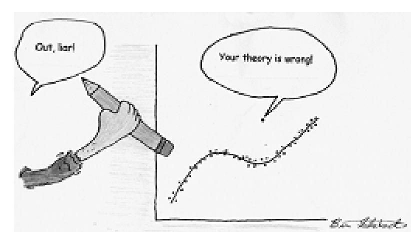
- Exploratory data analysis can be used to identify errors. Simple plots such as histograms or scatter plots of all variables (not only the response variable!!) can be used to look for weird data points – use common sense, e.g. negative body weights or animals died before they were born.

Slide 8

Outliers



- A value that is far from the others: it is an unusually large or an unusually small value compared to the others



Slide 9

What to do – delete or keep in the analysis?

1. Was the value entered into the computer correctly? If there was an error in data entry, fix it.
2. Is there a justification to exclude the value resulting from that analysis?
3. Is the outlier caused by "normal" variation? The observation/individual may be different from the others. This may be the most exciting finding in your data!

Slide 10

Outlier

- One way to identify univariate outliers is to convert all of the scores for a variable to standard scores.
- If the sample size is small (80 or fewer cases), a case is an outlier if its standard score is ±2.5 or beyond.
- If the sample size is larger than 80 cases, a case is an outlier if its standard score is ±3.0 or beyond

Slide 11

Example: **Grazing behavior of sheep**

- The researcher in this project is convinced that male sheep loop up more frequently while eating than female sheep (and also has a theory on why this should be the case). An undergraduate student spends many uncomfortable weeks in a hide near a field (so as not to disturb the sheep) and records the data for each observation on each of 3 male and 3 female sheep.

Slide 12

Research question :

Do male sheep look up more frequently while eating than female sheep?

- Variables:
- DURAT: duration of feeding time in minutes
- NLUPS: the number of lookups
- SEX: coded as 1 for female and 2 for male
- SHEEP: coded as 1 to 6
- OBSP: the number of the observation period from 1 to 20.

Slide 13

Initial examination of the data.

Variable: NLUPS

Moments

N	120	Sum Weights	120
Mean	7.85	Sum Observations	942
Std Deviation	2.46879687	Variance	6.09495798
Skewness	-0.0324914	Kurtosis	-0.4887654
Uncorrected SS	8120	Corrected SS	725.3
Coeff Variation	31.4496417	Std Error Mean	0.22536929

Basic Statistical Measures

Location		Variability	
Mean	7.850000	Std Deviation	2.46880
Median	7.000000	Variance	6.09496
Mode	7.000000	Range	12.00000
		Interquartile Range	4.00000

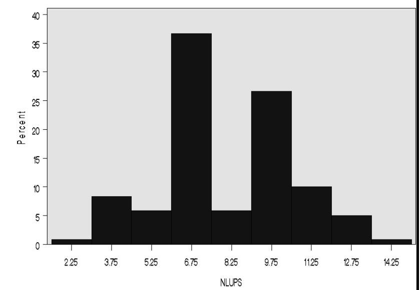
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
2	9	12	110
3	12	12	111
3	11	12	118
3	7	13	113
3	3	14	115

Slide 14

Response variable (Y) :

How are they distributed? Do not need to be normally distributed but a clear bimodal pattern or a strange value (outlier) might give a reason to be cautious.



Slide 15

2) Relations explanatory variables and the response variable.

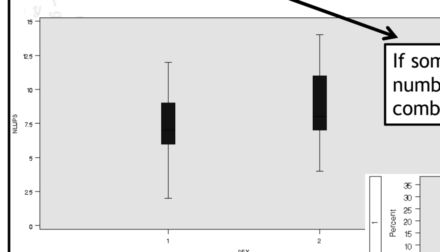
- Get a first, rough idea about the effect of the fixed effect classes.
- Get a first clue about the type of relationship between the regressors and the response variables.

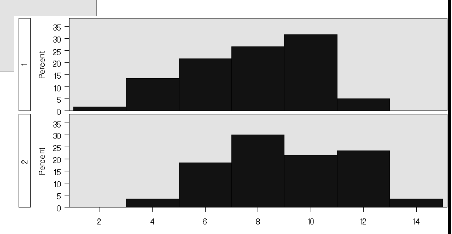
Slide 16

Analysis Variable : NLUPS

SEX	N Obs	Mean	Std Dev	N	N Miss	Minimum	Maximum	Median
1	60	7.2000000	2.4272465	60	0	2.0000000	12.0000000	7.0000000
2	60	8.5000000	2.3542245	60	0	4.0000000	14.0000000	8.0000000

If some classes contain only a very small number of observations we might want to combine classes

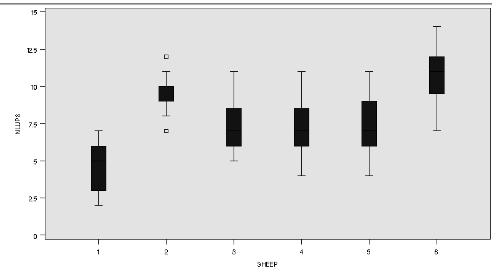




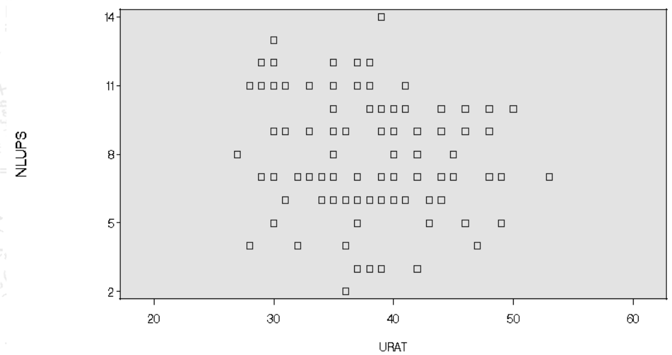
Slide 17

Analysis Variable : NLUPS

SHEEP	N Obs	Mean	Std Dev	N	N Miss	Minimum	Maximum	Median
1	20	4.8000000	1.6415654	20	0	2.0000000	7.0000000	5.0000000
2	20	9.5000000	1.1920791	20	0	7.0000000	12.0000000	10.0000000
3	20	7.3000000	1.5927467	20	0	5.0000000	11.0000000	7.0000000
4	20	7.2500000	1.7733406	20	0	4.0000000	11.0000000	7.0000000
5	20	7.5000000	1.7917942	20	0	4.0000000	11.0000000	7.0000000
6	20	10.7500000	1.6819475	20	0	7.0000000	14.0000000	11.0000000



Slide 18



Slide 19

3) Relations among explanatory variables.

Explanatory variables might (partly) explain the same variation in the response variable.

Confounding:

Two variables are confounded if they vary together in such a way that it is impossible to determine which variable is responsible for an observed effect.

Slide 20

Analysis Variable : URAT

SEX	N Obs	Mean	Std Dev	N	N Miss	Minimum	Maximum	Median
1	60	40.5833333	5.1561767	60	0	28.0000000	53.0000000	40.5000000
2	60	35.6000000	4.7271305	60	0	27.0000000	48.0000000	35.0000000

Analysis Variable : URAT

SHEEP	N Obs	Mean	Std Dev	N	N Miss	Minimum	Maximum	Median
1	20	36.8000000	3.8333715	20	0	28.0000000	43.0000000	37.0000000
2	20	40.7500000	4.3270873	20	0	35.0000000	50.0000000	41.0000000
3	20	44.2000000	4.5026308	20	0	37.0000000	53.0000000	44.0000000
4	20	38.1500000	5.0186494	20	0	30.0000000	48.0000000	39.0000000
5	20	34.8000000	4.3115512	20	0	27.0000000	46.0000000	35.0000000
6	20	33.8500000	3.8563004	20	0	28.0000000	40.0000000	33.5000000

Slide 21

- Experiment comparing two treatments for depression In case of a significant difference between treatment groups treatments were found, it is impossible to say if the effect is due to treatment or due to an age difference

		Treatment 1	Treatment 2
Age	Young	*	
	Old		*

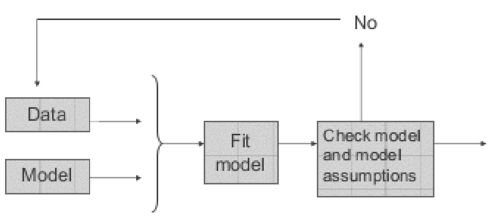
Slide 22

Choose the appropriate statistical model.

- Consideration:
- Duration is continues variable, likely there isn't a relationship between URAT & NLUP.
- Categorical variable are sex and sheep (sex).
- Existence of confounding between sheep and sex

Slide 23

Model building



Be prepared to try more than one analysis.

Model building is partly science and partly art.

Slide 24

The GLM Procedure

Class Level Information

Class	Levels	Values
SEX	2	1 2
SHEEP	6	1 2 3 4 5 6

Number of Observations Read 120
Number of Observations Used 120

18:31 Sunday, March 9, 2003

The GLM Procedure

Dependent Variable: NLUPS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	426.3794326	71.0632388	26.86	<.0001
Error	113	298.9205674	2.6453148		
Corrected Total	119	725.3000000			

Slide 25

R-Square	Coeff Var	Root MSE	NLUPS Mean
0.587866	20.71901	1.626442	7.850000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
URAT	1	1.9794326	1.9794326	0.75	0.3889
SHEEP(SEX)	4	375.6707731	93.9176933	35.50	<.0001
SEX	1	29.3200314	29.3200314	11.08	0.0012

18:31 Sunday, March 9, 2003

The GLM Procedure
Least Squares Means

SEX	NLUPS LSMEAN
1	7.27587825
2	8.42412175

Slide 26

The GLM Procedure

Class Level Information

Class	Levels	Values
SEX	2	1 2
SHEEP	6	1 2 3 4 5 6

Number of Observations Read 120
Number of Observations Used 120

19:05 Sunday, March 9, 2003

The GLM Procedure

Dependent Variable: NLUPS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	424.4000000	84.8800000	32.16	<.0001
Error	114	300.9000000	2.6394737		
Corrected Total	119	725.3000000			

Slide 27

R-Square	Coeff Var	Root MSE	NLUPS Mean
0.585137	20.69612	1.624646	7.850000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	50.7000000	50.7000000	19.21	<.0001
SHEEP(SEX)	4	373.7000000	93.4250000	35.40	<.0001

19:05 Sunday, March 9, 2003

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

SEX	NLUPS LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
1	7.20000000	<.0001
2	8.50000000	

Slide 28

Model Assumptions

- Independence (response variables y_i are independent)- this is a design issue
- Normality (response variables are normally distributed)
- Homoscedasticity (the response variables have the same variance)

Slide 29

- Gross violations of the assumptions may yield an unstable model with opposite conclusions.
- The standard summary statistics: t-; F- statistics and R^2 can not detect the departures from the underlying assumptions.
- Based on the study of the model residuals.

Slide 30

Best way to check assumptions

check the assumptions on the *random errors*

- They are **independent**
- They are **normally distributed**
- They have a **constant variance σ^2 for all settings of the independent variables (Homoscedasticity)**
- They have a **zero mean.**

If these assumptions are satisfied, we may use the normal density as the working approximation for the random component. So, the residuals are distributed as:

$\varepsilon_i \sim N(0,\sigma^2)$

Slide 31

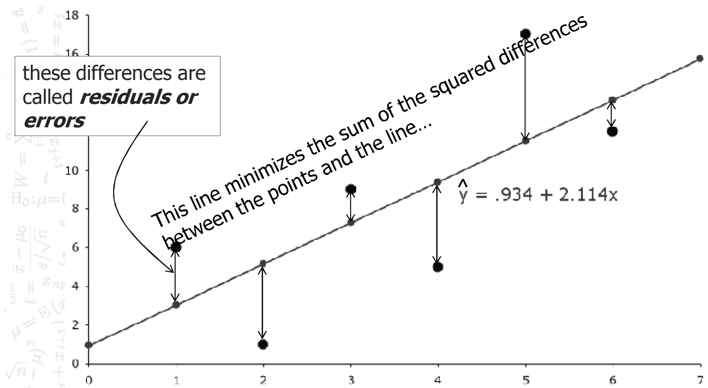
Residual Analysis

Definition of Residuals

- Residual:
  - The deviation between the data and the fit
    $e_i = y_i - \hat{y}_i, i = 1, \cdots, n$
  - A measure of the variability in the response variable not explained by the regression model.
  - The realized or observed values of the model errors.

Slide 32

Least Squares Line…



Slide 33

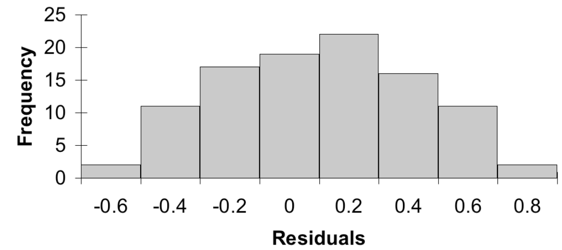
Normality

- The random errors can be regarded as a random sample from a $N(0,\sigma^2)$ distribution, so we can check this assumption by checking whether the residuals might have come from a normal distribution.
- We should look at the standardized residuals
- Options for looking at distribution:
  - Histogram, Stem and leaf plot, Normal plot of residuals

Slide 34

Nonnormality…

- We can take the residuals and put them into a histogram to visually check for normality…



- …we're looking for a bell shaped histogram with the mean close to zero [our aim "test for normality]. ✓

Slide 35

Normal Plot of Residuals

- A normal probability plot is found by plotting the residuals of the observed sample against the corresponding residuals of a standard normal distribution *N* (0,1)
  - If the plot shows a **straight line**, it is reasonable to assume that the observed sample comes from a normal distribution.
  - If the points deviate a lot from a straight line, there is evidence against the assumption that the random errors are an independent sample from a normal distribution.

Slide 36

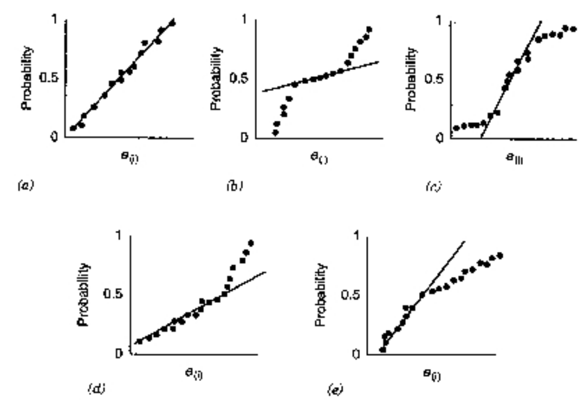


Figure 4.1 Normal probability plots: (a) ideal; (b) heavy-tailed distribution; (c) light-tailed distribution; (d) positive skew; (e) negative skew.

Slide 37

Plotting Residuals

- To check for **Homoscedasticity** (constant variance):
  - Produce a scatter plot of the standardized residuals against the fitted values.
  - Produce a scatter plot of the standardized residuals against each of the independent variables.
- If assumptions are satisfied, residuals should vary randomly around zero and the spread of the residuals should be about the same throughout the plot (no systematic patterns.)

Slide 38

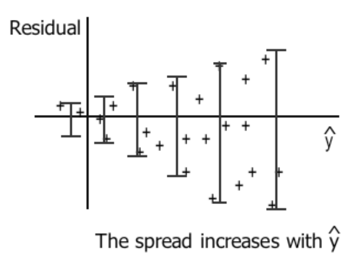
Homoscedasticity is probably violated if…

- The residuals seem to increase or decrease in average magnitude with the fitted values, it is an indication that the variance of the residuals is not constant.
- The points in the plot lie on a curve around zero, rather than fluctuating randomly.
- A few points in the plot lie a long way from the rest of the points.

Slide 39

Heteroscedasticity…

- When the requirement of a constant variance is violated, we have a condition of ***heteroscedasticity***.



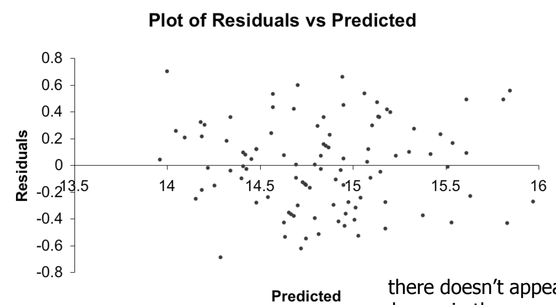
- We can diagnose heteroscedasticity by plotting the residual against the predicted y.

17.39

Slide 40

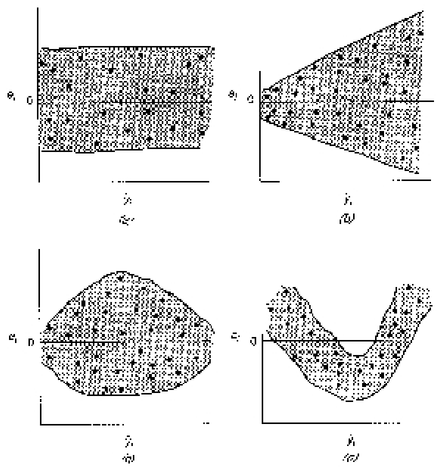
Heteroscedasticity…

- Here's the plot of the residual against the predicted value of **y**:



there doesn't appear to be a change in the ***spread*** of the plotted points, therefore no ***heteroscedasticity*** ✓

41



Slide 42

Plot of Residuals against the Fitted Values:

From Fig :

1. Fig a: Satisfactory
2. Fig b: Variance is an increase function of y
3. Fig c: Often occurs when y is a proportion between 0 and 1.
4. Fig d: Indicate nonlinearity.

- For 2. and 3., use suitable transformations to either the regressor or the response variable or use the method of weighted LS.
- For 4., except the above two methods, the other regressors are needed in the model.

43

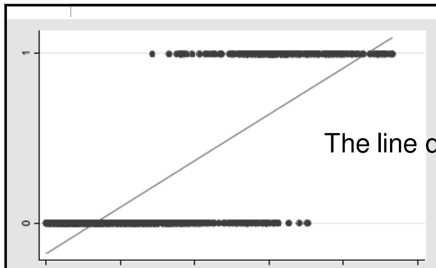
Nonlinear Estimation

- Until now Y, the dependent variable, was continuous.
- Independent variables could be dichotomous (dummy variables).
- We’ll start our exploration of non-linear estimation with dichotomous Y vars.
- These arise in many science problems:
  - Survival (0 and 1)
  - Calving ease (1 , 2 , 3 , 4)

44

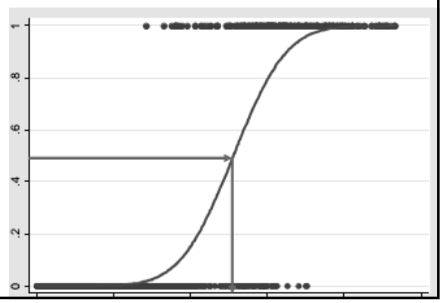
Nonlinear models

In these examples, the dependent variables are not continuous, and classical regression or analysis of variance may not be appropriate because assumptions such as homogeneity of variance and linearity are often not satisfied. Further, these variables do not have normal distributions and *F or t tests are* not valid.



The line doesn't fit the data very well.

So we need a link function F(Y) that takes a dichotomous Y and gives us a continuous, real-valued



The Structure of Generalized Linear Models

Slide 46

Generalized linear models are models in which independent variables explain a function of the mean of a dependent variable. This is in contrast to classical linear models in which the independent variables explain the dependent variable or its mean directly. Which function is applicable depends on the distribution of the dependent variable.

- generalized linear models (GLMs) extend the range of application of linear statistical models by accommodating response variables with non-normal conditional distributions.
- Except for the error, the right-hand side of a generalized linear model is essentially the same as for a linear model.

A generalized linear model consists of <u>three</u> components:

Slide 48

1. <u>***A random component***</u>*, specifying the conditional distribution of the* response variable, y_i ,given the explanatory variables.

- Traditionally, the random component is a member of an "exponential family" (the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions) but generalized linear models have been extended beyond the exponential families.

Slide 49

Normal distribution—a family of distributions, each member of which can be defined by the mean and variance—many physical phenomena can be approximated well by the normal distribution.

Binomial distribution—probability distribution of # of successes in a sequence of Bermoulli trials (where outcomes fall into one of two categories—i.e., ”occurred” and “did not occur”. Note that in large samples, if the dependent variable is not too skewed, then the normal distribution approximates the binomial distribution.

Slide 50

Poisson Distribution

- expresses the probability of a # of events occurring in a fixed period of time, if the events occur with a known average rate, and independently of the time since the last event.
- Poisson distributions are often used in modeling count data. Poisson random variables take on non-negative integer values, 0, 1, 2, … .

2. A linear function of the regressors (linear predictor)

$$\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} = \mathbf{x}_i'\boldsymbol{\beta}$$

on which the expected value μ_i of y_i depends.

- The X’s may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, contrasts generated from factors, interaction regressors, etc.

Slide 52

3. ***An invertible link function***, $g(\mu_i) = \eta_i$

which transforms the expectation of the response to the linear predictor.

- The inverse of the link function is sometimes called the *mean function:*

$$g^{-1}(\eta_i) = \mu_i$$

Standard link functions and their inverses:

Slide 53

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

The logit, probit, and complementary-log-log links are for *binomial data*, where Y_i represents the observed proportion and μ_i the expected proportion of "successes" in n_i binomial trials — that is, μ_i is the probability of a success.