STATA
Note 5

One sample binomial data
Confidence interval for proportion
Unpaired binomial data: 2 x 2 tables
Paired binomial data

One sample binomial data (Test for $\pi = \pi_0$)

One can in general be in one of two situations:

A. Data are available in the data file with one line per person (individual-level data).
   Menu: Statistics ▶ Parametric tests ▶ 1-sample test of proportion
   Insert data variable (must be a 1/0 variable) and π_0.
   Confidence level may be given in percentage.
   Syntax: `bintest` *`datavariable`* `=` π_0`, level(95)`

B. Data are summarised by the total number of observations (n) and the total number of successes (k) (summarised data). Can only use syntax.
   Syntax: `bitesti n k` π_0

Example:

Syntax: `bitesti 88 35 0.5`

Output:

```
        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       88         35           44       0.50000      0.39773

  Pr(k >= 35)            = 0.978895  (one-sided test)
  Pr(k <= 35)            = 0.034673  (one-sided test)
  Pr(k <= 35 or k >= 53) = 0.069346  (two-sided test)
```

Exact confidence interval for proportion

Again we distinguish between the two situations.
Situation A.
Syntax: `ci` *`datavariable`*`, b`

Situation B.
Menu:
   Calculator ▶ Binomial confidence interval
   Insert number of observations (n), number of successes (k).
   Confidence level may be given in percentage.

Syntax: `Cii  n k`

Example:

Syntax: `cii 88 35, level(95)`

Output:

```
                                                   -- Binomial Exact --
Variable |      Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
         |       88    .3977273    .0521733        .2948726    .5076575
```

General on tables

One can in general be in one of two situations:
A. Data are available in the data file with one line per person (individual-level data).
B. Data are already tabulated; e.g. one has the table on paper (summarised data).

Situation A is straightforward:
Syntax: `tabulate` *`rowvariable columnvariable`*

Situation B is often dealt with in the following way:
Data:

	Status		
Diagnosis	Alive	Dead	Total
1	1600	40	1640
2	1780	20	1800
Total	3380	60	3440

1. Generate two variables, `Diagnosi` and `Status`, and four observations, i.e. four lines with the values specifying the table (see item 2).
2. Generate a third variable `counts` containing the counts in the table:
```
DIAGNOSI STATUS     COUNTS
1        Dead         40
1        Alive      1600
2        Dead         20
2        Alive      1780
```
3. Syntax: `tabulate` *`diagnos status`* `[fweight=`*`counts`*`]`

Output:

```
           |        status
   diagnos |     alive       dead |     Total
-----------+----------------------+----------
         1 |      1600         40 |      1640
         2 |      1780         20 |      1800
-----------+----------------------+----------
     Total |      3380         60 |      3440
```

Unpaired binomial data: 2 x 2 tables

Both in situation A and situation B proportions relative to **row or column total** may be calculated by adding the word `row` or `column` in the syntax-line, as for example:
Syntax: `tabulate` *`diagnos status`* `[fweight=`*`counts`*`], row`

A **chi-square test** is made adding the word `chi2` after the comma in the syntax-line.
Further, if the table is sparse then **Fishers exact test** should be used. This is done typing `exact` after the comma in the syntax-line.

Stata can estimate the **risk difference**, **relative risks** (also called risk ratio) and **odds ratios**; all with confidence interval. But tables designed for the specific case should be used - Stata codes are in the following listed in parenthesis.
The risk difference and relative risks can be found using a cohort study (`cs`) or a matched case-control study (`mcc`). Odds ratios are typically estimated using case-control (`cc`) or matched case-control studies (`mcc`).Further help may be found in the Stata on-line help manual under `epitab`.
Such epidemiological tables are usually denoted as

	exposed	unexposed
cases	a	b
Non-cases	c	d

If you do not have data in the original form but data are already tabulated (i.e., *summarised data*), then the statistical analysis is made using the immediate commands in Stata. That is, an `i` is added in the syntax and data are in compact form typed as arguments of the command. For example, if the table on paper is in a form similar to the above table, then estimates of the risk difference can easily be found using the syntax of a cohort study.
Syntax: `csi a b c d`

Note: Epidemiologists use terms as exposed/unexposed and cases/non-cases, so output is presented in tables using these labels. However, the procedures (`epitab`) may be used for general analyses. One should just remember that exposed/unexposed divide data according to the two groups to be compared (for example, two treatments groups or men and women) and cases/non-cases divide data according to the outcome of interest, e.g. death and non-death (=survival).

Note: If you have individual-level data (i.e., the data file consists of one line of data per person) then the procedures are all used specifying the table command followed by case- and exposure-variables, e.g.
Syntax: `cs` *`case-variable exposure-variable`*

Note: The commands assume the subject to be a case if the *case-variable* is 1 and a non-case if the *case-variable* is 0. Further, the subject is considered as exposed if the *exposure-variable* is nonzero and not missing, to be unexposed if the *exposure-variable* is zero and will ignore a subject if the *exposure-variable* is missing.

Example: Comparison of two treatments.

We want to assess the value of a new treatment A in comparison with the old treatment B (exposed = treatment A and unexposed = treatment B). We want to compare the risk of dying, hence cases = deaths and non-cases = survived. Data are tabulated as

	Treatment A	Treatment B
Death	41	64
Survival	216	180

Syntax: `csi 41 64 216 180`

Output:

```
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        41          64  |        105
        Noncases |       216         180  |        396
-----------------+------------------------+----------
           Total |       257         244  |        501
                 |                        |
            Risk |  .1595331    .2622951  |  .2095808
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.102762        |  -.173829    -.031695
      Risk ratio |        .6082198        |  .4282837    .8637531
 Prev. frac. ex. |        .3917802        |  .1362469    .5717163
 Prev. frac. pop |        .2009731        |
                 +-----------------------------------------------
                           chi2(1) =     7.98  Pr>chi2 = 0.0047
```

Paired binomial data (McNemar's test)

In Stata McNemar's chi-squared test requires data to be represented as matched case-control data. Again if data are already tabulated the McNemar's chi-squared test is easily performed using the syntax.
Syntax: `mcci a b c d`

Example: Severe colds

Syntax: `mcci 212 144 256 707`

Output:

```
                 | Controls               |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
         Exposed |       212         144  |        356
       Unexposed |       256         707  |        963
-----------------+------------------------+----------
           Total |       468         851  |       1319

McNemar’s chi2(1) =     31.36        Pr>chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases       .2699014
        Controls    .3548143     [95% conf. interval]
                    ---------    --------------------
        difference -.0849128     -.1150345  -.0547911
        ratio       .7606838      .6910295    .837359
        rel. diff. -.1316099       -.18061  -.0826098

        odds ratio     .5625      .4553956     .69259   (exact)
```

Note:
Stata do not use the continuity-corrected version of McNemar’s test. Therefore minor differences may be found comparing with the notes for the “Postgraduate course in Biostatistics” or other statistical-programs.

The estimated proportions are

- Cases: the estimated risk of being exposed among cases.
- Controls: the estimated risk of being exposed among controls.
- Rel. diff: the estimated risk difference of being exposed relative to the risk of being unexposed among controls.
- The rest of the results are ‘self-explained’.