```
      name:  <unnamed>
       log:  /Users/eric/Work/MB/ATCR/200/Labs/lab 3.smcl
  log type:  smcl
 opened on:   9 Jul 2018, 04:54:55

  1 .
  2 . ********************************************************************************
  3 . * Binomial distribution.
  4 .
  5 . * The probability of being selected by random digit dialing for a political survey in your area code is 0.025. The
    >  surveys are conducted randomly each month for one year.  What is the probability that you will not be selected in
    >  the entire year (i.e. the probability you will not be selected in any of 12 months)?  Use the binomial formula.
  6 .
  7 . * We will do this two ways using built-in three built-in Stata functions.  Briefly check the help files for these
    > functions
  8 .
  9 . help comb

 10 .
 11 . help binomialtail

 12 .
 13 . help binomialp

 14 .
 15 . * Now we implement the binomial formula directly using comb(N,k), which evaluates the combinatoric (N choose k), a
    > nd the exponentiation operator ^.
 16 .
 17 . display as result 1-comb(12,0)*0.025^0*(1-.025)^12
  .26200165

 18 .
 19 . * The as result option in the display command bolds the output. An easier way to do this is to use the binomialtai
    > l function.
 20 .
 21 . di as result binomialtail(12,1,.025)
  .26200165

 22 .
 23 . * Now we calculate the probability of being chosen once or twice, first using the binomial forumla
 24 .
 25 . di as result comb(12,1)*0.025^1*(1-.025)^11+comb(12,2)*0.025^2*(1-.025)^10
  .25910001

 26 .
 27 . * and now using the binomialp function
 28 . di as result binomialp(12,1,.025)+binomialp(12,2,.025)
  .25910001

 29 .
 30 . * Now we use binomialtail to get the probability of being selected 2 times or fewer:
 31 . di as result 1-binomialtail(12,3,.025)
  .99709836

 32 .
 33 . * Now we will get summary statistics and plot the binomial distribution with first 10, then 100, trials and varyin
    > g probabilities of success.  Using loops, we'll first make a dataset with 11 or 101 observations, one for each pos
    > sible number of successes, then calculate the corresponding probability.
 34 .
 35 . foreach n in 10 100 {
   2.         clear
   3.         set obs `=`n'+1' // note syntax to plug in the sample size + 1
   4.         gen x = _n-1 // _n is the observation number
   5.         foreach p in .5 .18 .82 {
   6.
 36 .                 * calculate mean and SD
 37 .                 local mean = round(`n'*`p',.01)
```

```
      7.                  local sd = round(sqrt(`n'*`p'*(1-`p')),.01)
      8.                  di as result "N = `n', P = `p', mean = `mean', SD = `sd'"
      9.
 38 .                  * calculate probability of each number of successes
 39 .                  local pct = `p'*100 // variable and plot names can't include "."
     10.                  gen p`pct' = binomialp(`n',x,`p')
     11.                  * plot the distribution
 40 .                  twoway (bar p`pct' x, fcolor(blue)), ///
    >                          title("N = `n', P = `pct'%") caption("mean = `mean', SD = `sd'") ///
    >                          name(b`n'_`pct', replace)
     12.                  }
     13.         }
    number of observations (_N) was 0, now 11
    N = 10, P = .5, mean = 5, SD = 1.58
    N = 10, P = .18, mean = 1.8, SD = 1.21
    N = 10, P = .82, mean = 8.199999999999999, SD = 1.21
    number of observations (_N) was 0, now 101
    N = 100, P = .5, mean = 50, SD = 5
    N = 100, P = .18, mean = 18, SD = 3.84
    N = 100, P = .82, mean = 82, SD = 3.84

 41 . * Note that Stata has trouble rounding one of the means; this comes from the conversion back from binary to decima
    > l notation.  Now combine the plots to make them easier to compare.
 42 .
 43 . graph combine b10_50 b10_18 b10_82 b100_50 b100_18 b100_82, rows(2)

 44 .
 45 . * Are the plots consistent with the calculated means and SDs?
 46 .
 47 . * According to the CDC, 20% of all adult men in the U.S. smoke cigarettes.  If you selected repeated samples of si
    > ze 12 from all adult males in the U.S., what would be the mean number of men who do not smoke? What would be the m
    > ean number of men who smoke? What would be the standard deviation?
 48 .
 49 . di as result "Mean # non-smokers = " round(12*.8,.01)
    Mean # non-smokers = 9.6

 50 . di as result "Mean # smokers = " round(12*.2,.01)
    Mean # smokers = 2.4

 51 . di as result "SD of distribution = " round(sqrt(12*.8*.2),.01)
    SD of distribution = 1.39

 52 .
 53 . * Suppose you select a sample of 12 men and find that 6 of them smoke.  Assuming .20 is the true probability of sm
    > oking among men, what is the probability that you would have a sample this extreme?
 54 .
 55 . di as result binomialtail(12,6,.2)
    .01940528

 56 .
 57 . ********************************************************************************
 58 . * Hypergeometric distribution
 59 .
 60 . * A binomial outcome k can be thought of as the number of successes in a sample of size n, sampled with replacemen
    > t from a population of size N, in which the proportion of successes K/N=p. The corresponding hypergeometric outcom
    > e k is the number of successes in a sample of size n, in this case sampled without replacement from the same popul
    > ation.  Commonly we assume that the population is so large that sampling with or without replacement is of no prac
    > tical importance.  But what if N is smalle?
 61 .
 62 . * Dr. Allen gives this New York City-based example:  Having bought a bag of 30 roasted chestnuts, you walk home ea
    > ting them with gusto, and arrive home having eaten 20.  In opening the remaining 10, you find that 7 contain worms
    > .  What is the probability that none if the first 20 contains worms?
 63 .
 64 . * First check the help for the hypergeometric probability functions.
 65 .
 66 . help hypergeometricp

 67 .
```

```
 68 . help hypergeometric

 69 .
 70 . * We can rephrase the problem as follows:  In a sample of 20 sampled without replacement from a population of size
    >  30 including 7 "successes", what is the probability of drawing 0 successes?
 71 .
 72 . di as result "probability you recently ate no wormy chestnuts = " ///
    >         round(hypergeometric(30,7,20,0),.000001)
    probability you recently ate no wormy chestnuts = .000059

 73 .
 74 . * We can also calculate this from the formula
 75 . di as result "probability you recently ate no wormy chestnuts = " ///
    >         round(comb(7,0)*comb(23,20)/comb(30,20),.000001)
    probability you recently ate no wormy chestnuts = .000059

 76 .
 77 . * Note that we could have used either hypergeometric function in this case.
 78 .
 79 . ********************************************************************************
 80 . * Normal distribution
 81 .
 82 . * First, familiarize yourself with the Stata normal cumulative probability  and inverse cumulative probability fun
    > ctions
 83 .
 84 . help normal

 85 .
 86 . help invnormal

 87 .
 88 . * The lengths of adult men's feet are approximately normal, with mean 11 inches, and standard deviation 1.5 inches
    > .
 89 .
 90 . * What is the probability of a foot length of more than 13 inches? Because the argument to the normal function is
    > Z-score (i.e., the observed value minus the mean, divided by the SD), we calculate that first.  Using a local vari
    > able avoids typing in a calculated number, which is prone to error. Before running this next bit of code, try to p
    > ut in words what it is doing.
 91 .
 92 . local z = (13-11)/1.5

 93 . di as result "probability of foot length >=13 inches: " ///
    >         round(1-normal(`z'),.001)
    probability of foot length >=13 inches: .091

 94 .
 95 . * We could also do this in a single step, because the normal function can calculate its argument on the fly.  (Sta
    > ta is inconsistent about this.)
 96 .
 97 . di as result "probability of foot length >=13 inches: " ///
    >         round(1-normal((13-11)/1.5),.001)
    probability of foot length >=13 inches: .091

 98 .
 99 . * What is the probability of a foot length between 10 and 12 inches?
100 .
101 . local z10 = (10-11)/1.5

102 . local z12 = (12-11)/1.5

103 . local p1012 = round(normal(`z12')-normal(`z10'),.01)

104 . di as result "probability of foot length 10-12 inches: `p1012'"
    probability of foot length 10-12 inches: .5

105 .
106 . * This could also have been done in a single step; try that if you like.
107 .
```

```
108 . * What are the foot lengths at the 25th percentile and 75th percentile? The argument of the invnormal function is
    > a probability and the output is a Z-score, from which we calculate the observed value using the mean and SD.
109 .
110 . di as result "25th pctile = " round(11+invnormal(.25)*1.5,.01) ///
    >         "  inches"  _n "75th pctile = " round(11+invnormal(.75)*1.5,.01) " inches"
    25th pctile = 9.99  inches
    75th pctile = 12.01 inches

111 .
112 . * The probability of 0.04 that a randomly chosen adult male foot length will be less than how many inches?
113 .
114 . di as result "4th pctile = " round(11+invnormal(.04)*1.5,.01) "inches"
    4th pctile = 8.37inches

115 .
116 . * Stata had trouble with the rounding again.
117 .
118 . * We could also have done this using simulation, which is commonly used when a distribution isn't well understood
    > (in contrast to the Normal distribution).  To get reasonably reliable estimates, we'll use a large simulated sampl
    > e of 100,000, and also set the seed for the random number generator, to make the results reproducible.
119 .
120 . clear

121 . set obs 100000
    number of observations (_N) was 0, now 100,000

122 . set seed 200

123 . gen fl = rnormal(11,1.5)

124 . label var fl "foot length"

125 .
126 . * Following lab 1, we'll use the ordering of the elements of the recode command to ensure that any values of 13 ar
    > e counted in the >= 13 category.
127 .
128 . recode fl (13/max=1 ">=13 inches") (min/13=0 "<13 inches"), gen(fl_13up)
    (100000 differences between fl and fl_13up)

129 . label var fl_13up "foot length"

130 . tab fl_13up
```

foot length	Freq.	Percent	Cum.
<13 inches	**90,812**	**90.81**	**90.81**
>=13 inches	**9,188**	**9.19**	**100.00**
Total	**100,000**	**100.00**	

```
131 .
132 . recode fl (10/12=1 "10-12 inches") (min/10 12/max=0 "<10, >12 inches"), gen(fl_1012)
    (100000 differences between fl and fl_1012)

133 . label var fl_1012 "foot length"

134 . tab fl_1012
```

foot length	Freq.	Percent	Cum.
<10, >12 inches	**50,779**	**50.78**	**50.78**
10-12 inches	**49,221**	**49.22**	**100.00**
Total	**100,000**	**100.00**	

```
135 .
136 . * The result is close to the theoretical value of 9.1%.  We can also get the quantiles of the distribution as loca
    > l variables, then use the command return list to see the results.
```

```
137 .
138 . _pctile fl, percentiles(4 25 75)

139 . return list

    scalars:
                     r(r1) =  8.369423389434814
                     r(r2) =  9.983631134033203
                     r(r3) =  12.02052927017212

140 .
141 . * Results are again close to the theoretical values. To get a sense of simulation error, try redoing this with a s
    > maller simulated sample.
142 .
143 . ********************************************************************************
144 . * Poisson and negative binomial distributions
145 .
146 . * Count outcomes are commonly encountered in epidemiology.  Examples include the number of cases of incident disea
    > se in a given number of person-years of follow-up, the number of binge drinking episodes reported between study vi
    > sits, and so on.
147 .
148 . * Check out the four Poisson probability functions:
149 .
150 . help poissonp

151 . help poisson

152 . help poissontail

153 . help invpoisson

154 . help invpoissontail

155 .
156 . * Suppose you accrue 10,545 person-years of follow-up in your cohort, and the population incidence of myocardial i
    > nfarction is 2.5% (i.e., 2.5 cases per 100 person-years).  Assuming that the distribution is Poisson, which is rea
    > sonable in this case, what is the expected number of cases?  We'll save the expected number of cases as a global v
    > ariable, which will be remembered even if you run the commands below one at at time; note that we call global vari
    > ables using a different syntax.
157 .
158 . global mean = 10545*.025

159 . di as result "Expected number of cases = $En"
    Expected number of cases =

160 .
161 . * What is the probability of observing at least 300 MI cases?
162 .
163 . di as result "probability of observing >=300 MI cases: " ///
    >         round(poissontail($mean,300),.001)
    probability of observing >=300 MI cases: .015

164 .
165 . * What is the chance of observing between 250 and 275 cases?
166 .
167 . di as result "probability of observing 250-275 MI cases: " ///
    >         round(poisson($mean,275)-poisson($mean,249),.001)
    probability of observing 250-275 MI cases: .576

168 .
169 . * What is the IQR of the distribution?
170 .
171 . local q25 = round(invpoisson($mean, .25),.1)

172 . local q75 = round(invpoisson($mean, .75),.1)

173 . di as result ///
    >         "IQR of Poisson distribution with mean `=round($mean,.1)': `q25'-`q75'"
```

```
    IQR of Poisson distribution with mean 263.6: 274.8-252.9

174 .
175 . * Although the Poisson model might be adequate for the number of MIs, the number of binge drinking episodes might
    > be "over-dispersed", with variance greater than its mean.  The negative binomial distribution, with parameters n,
    > k, and p, gives the probabililty of observing k failures before n successes are observed, where p is the probabili
    > ty of success.
176 .
177 . * That of course is a counter-intuitive way to think of the distribution. Consider that the mean of the distributi
    > on of k is n(1-p)/p and its variance is n(1-p)/p^2.  A little algebra shows that p is the ratio of the mean to the
    >  variance, and n is given by mean^2/(variance-mean). Consider a negative binomial distribution with mean 2 and var
    > iance 5 (so it is over-dispersed, relative to the Poisson model).  We can check to see that these formulas are cor
    > rect.
178 .
179 . local m = 2

180 . local v = 5

181 . local p = `m'/`v'

182 . local n = `m'^2/(`v'-`m')

183 . clear

184 . set obs 100000
    number of observations (_N) was 0, now 100,000

185 . gen y = rnbinomial(`n',`p')

186 . tabstat y, s(mean var)

        variable |      mean  variance
    -------------+--------------------
               y |   1.99457   4.95259
    ----------------------------------

187 . clear

188 .
189 . * Now familiarize with the negative binomial probability functions
190 .
191 . help nbinomialp

192 . help nbinomial

193 . help nbinomialtail

194 . help invbinomial

195 . help invbinomialtail

196 .
197 . * Use these functions to calculate the probability of observing exactly 10 failures before the 5th success, if the
    >  probability of success is .4. Also calculate the mean and variance of the distribution.
198 .
199 . di as result "Probability of 10 failures before 5th success = " ///
    >         round(nbinomialp(5,10,.4),.01)
    Probability of 10 failures before 5th success = .06

200 . di as result "Mean = " 5*.6/.4 "  Variance = " 5*.6/.4^2
    Mean = 7.5  Variance = 18.75

201 .
202 . * Now calculate the probability of observing 5-10 failures before the 7th success, with success probabiity .6.
203 .
204 . di as result  "Probability of 5-10 failures before 7th success = " ///
    >         round(nbinomial(7,10,.6)-nbinomial(7,4,.6),.01)
    Probability of 5-10 failures before 7th success = .43
```

```
205 .
206 . ********************************************************************************
207 . * Exponential failure time distribution
208 .
209 . * The exponential distribution describes time to an endpoint if the failure rate is a constant, and is closely rel
    > ated to the Poisson distribution, as explained in class. Stata parameterizes the distribution in terms of the its
    > mean or scale, which is the inverse of the failure rate per unit time.  Have a look at the stat functions for this
    >  distribution:
210 .
211 . help exponential

212 . help invexponential

213 .
214 . * Suppose the failure rate of 10% per year, so the mean failure time is 10 years.  What is the probability of fail
    > ing within 2, 5, and 10 time units?
215 .
216 . forvalues t = 1/10 {
      2.         local p_`t' = round(exponential(10,`t'),.01)
      3.         di as result "Probability of failure within `t' years: `p_`t''"
      4.         }
    Probability of failure within 1 years: .1
    Probability of failure within 2 years: .18
    Probability of failure within 3 years: .26
    Probability of failure within 4 years: .33
    Probability of failure within 5 years: .39
    Probability of failure within 6 years: .45
    Probability of failure within 7 years: .5
    Probability of failure within 8 years: .55
    Probability of failure within 9 years: .59
    Probability of failure within 10 years: .63

217 .
218 . * Note that the cumulative risk is not a linear function of time.  Why?
219 .
220 . * What is the probability of surviving more than 8 years?  Note that the exponentialtail function is not what we w
    > ant for this purpose.
221 .
222 . di as result "Probability of surviving >8 years: " 1-round(exponential(10,8),.01)
    Probability of surviving >8 years: .45

223 .
224 . * Finally, get the IQR of the distribution
225 . local q25 = round(invexponential(10, .25),.01)

226 . local q75 = round(invexponential(10, .75),.01)

227 . di as result ///
    >         "IQR of exponential distribution with mean 10: `q25'-`q75' years"
    IQR of exponential distribution with mean 10: 2.88-13.86 years

228 .
229 . log close
          name:  <unnamed>
           log:  /Users/eric/Work/MB/ATCR/200/Labs/lab 3.smcl
      log type:  smcl
     closed on:   9 Jul 2018, 04:55:05
```