```
Stat 343, Mon 30-Nov-2020 -- Mon 30-Nov-2020
Probability and Statistics
Fall 2020


-------------------------------
Monday, November 30th 2020
-------------------------------
Wk 14, Mo
Topic:: Permutation testing with two binary categorical variables

Work with Fisher's convictions-twins data (from Example 2.7.4)
 twinDat <- data.frame(
  twin = rep(c("Di", "Mono"), times = c(17, 13)),
  conviction = rep(c("No", "Yes", "No", "Yes"), times = c(15, 2, 3, 10))
 )

Table:
 tally(conviction ~ twin, data=twinDat)

Q: What if the two variables were independent?
   Explore by permuting the values of one variable
     Like taking slips of paper, one per case, both values written on it
     Then cut the slips separating the values
       Place one half-slip in "conviction" bag, other half in "twin" bag
     Draw randomly one half-slip from each bag
       "tape" the half-slips together, making this a "case"
       draws are without replacement, so draw until both bags are empty

   In R (with mosaic package)
     shuffle( twinDat$twin )    % shuffles order of values in twin column
     tally(conviction ~ shuffle(twin), data=twinDat)    % for table

Produce several tables with permuted twin column, and note
 - marginal totals do not change
 - really only one degree of freedom: any one cell's value dictates all cells
 - connection to rhyper()
    compare cell [1,1] with result of rhyper(1, 17, 13, 18) over many trials:

randomizedStats <- do(5000)*tally(conviction~shuffle(twin),data=twinDat)[1,1]
gf_histogram(~result, data=randomizedStats, breaks=0:15, color="black") %>%
```

```
  gf_histogram(~rhyper(5000, 17, 13, 18), breaks=0:15, fill=~"red", color="red")
  or   gf_dist("hyper", params=list(17, 13, 18))

    what to compare cell [2,1] with? dhyper(1, ?, ?, ?):

Fit model to randomizedStats$result
 - look normal?
    gf_dhistogram(~result, data=randomizedStats, bins=10) %>%
        gf_fitdistr(dist="dnorm")
 - parameters of this normal distribution?
    favstats(~result, data=randomizedStats)
    gf_dhistogram(~result, data=randomizedStats, bins=10) %>%
        gf_dist("dnorm", params=list(mean=?, sd=?))

Compare cumulative probabilities:
 - from permutation statistics
    cumsum(tally(~result, randomizedStats)) / 5000
 - from pnorm() using parameters
    pnorm(6:15, mean=?, sd=?)
 - arrange side-by-side
    cbind(
        cumsum(tally(~result, randomizedStats)) / 5000,
        pnorm(6:15, mean=?, sd=?)
        )
    Not spectacular results
 - continuity correction?
    cbind(
        cumsum(tally(~result, randomizedStats)) / 5000,
        pnorm(6:15 + .5, mean=?, sd=?)
        )
    quite respectable results

Question:
 - What does rbinom(x, n, p) do?
 - How can its results be simulated?
```