1. *(30 points: 5,10,5,5,5)*
I have a box with 1 red balls, 3 blue balls, and 5 green balls.

(a) I draw 4 times WITHOUT replacement. What is the chance that I get exactly two green balls?

Version 1: $\binom{4}{2}\left(\frac{5}{9}\right)\left(\frac{4}{8}\right)\left(\frac{4}{7}\right)\left(\frac{3}{6}\right)$ $\binom{4}{2}\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)\left(\frac{5}{7}\right)\left(\frac{4}{6}\right)$ $\frac{\binom{5}{2}\binom{4}{2}}{\binom{9}{4}}$ Version 2: $\binom{4}{2}\frac{3}{9}\frac{2}{8}\frac{6}{7}\frac{5}{6}$

(b) I draw 4 times WITHOUT replacement. What is the chance that I get exactly 1 red or exactly two green balls?

version1: $\binom{4}{1}\left(\frac{1}{9}\right)(\cdot 1)^3$ + part a $- \frac{4!}{2!1!1!}\left(\frac{1}{9}\right)\left(\frac{5}{8}\right)\left(\frac{4}{7}\right)\left(\frac{3}{6}\right)$

version2: $\binom{4}{1}\left(\frac{2}{9}\right)\left(\frac{7}{8}\right)\left(\frac{6}{7}\right)\left(\frac{5}{6}\right)$ + part a $- \frac{4}{2!1!1!}\left(\frac{2}{9}\right)\left(\frac{3}{8}\right)\left(\frac{2}{7}\right)\left(\frac{4}{6}\right)$

(c) Consider two events A and B, draws are with replacement.
(A) A: at least 4 greens in 9 draws.
B: at least 19 greens in 36 draws.
Which chance is bigger? If possible, explain in terms of the Law of Averages. If not possible, explain why not.
it's 1 green below the expected value / want small chance error (CE) for abs. term.
chance error ↑ w/ sample size for abs. term.

(d) Consider two events A and B, draws are with replacement.
A: at least 4 greens in 9 draws.
(B) B: at least 16 greens in 36 draws.
Which chance is bigger? If possible, explain in terms of the Law of Averages. If not possible, explain why not.
- Small chance error for % term (choice B)
- CE. ↓ for % term

(e) Consider two events A and B, draws are with replacement.
(A) A: exactly 5 greens in 9 draws.
B: exactly 20 greens in 36 draws.
Which chance is bigger? If possible, explain in terms of the Law of Averages. If not possible, explain why not.
CE ↑ in abs. term w/ sample size.

2. *(15 points:5,5,5)*
Suppose we do the following simulation having to do with event A for part c) on the previous page. The box has 1 red balls, 3 blue balls, and 5 green balls.
Event A is that for 9 draws with replacement, we draw at least 4 greens.

(a) Each simulation should be 9 draws from the box with replacement, and for each simulation you should check to see if event A happens. Write code to do 10,000 replications of this simulation and compute the proportion of times that event A happens.

```
mean(rbinom(10000,9,5/9)>=4)
```

or

```
mean(replicate(10000, sum(sample(0:1, 9, T, prob=c(4/9,5/9)))>=4))
```

(b) True or false, and explain briefly: If you do this simulation in R, the proportion of times that event A happens should be quite close to the chance of event A.

True, long run proportion should be close to probability

(c) Suppose that for this simulation you create a vector of length 10,000, and each entry in the vector is the number of green balls drawn for that set of draws. True or false and explain briefly: A histogram for the numbers in this vector should look very much like the normal curve.

False, histogram close to probability histogram

# of draws small $9<25$

3. *(30 points: 10,5,15)*

Income $1000s	Number of households
30-50	30
50-80	120
80-100	60
100-150	75
150-200	15

This version:

%	width	height
10	20	0.5
40	30	$1.\bar{3}$
20	20	1
25	50	0.5
5	50	0.1

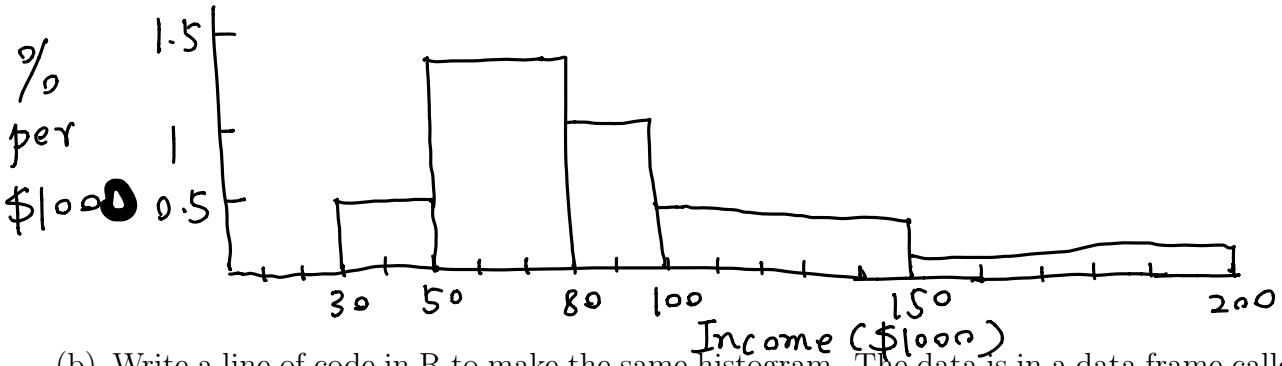
The other version:

%	width	height
10	20	0.5
35	20	1.75
25	30	0.83
25	50	0.5
5	50	0.1

The data above is for 300 household incomes in a particular neighborhood. You may assume that the incomes are evenly distributed within each interval and that the incomes are continuous.

(a) Draw a histogram including axes and density scale.

This version:



(b) Write a line of code in R to make the same histogram. The data is in a data frame called `data` and has a single column called `income`. Assume that the package `ggplot2` has been attached (so you can use any functions in that package). The histogram should use the density scale (total area 1), have the same class intervals as in the original dataset, and you should make the bars red with blue outlines.

This version:

```
ggplot(data, aes(x=income, y=..density..)) + geom_histogram(
breaks=c(30,50,80,100,150,200), col="blue", fill="red")
```

The other version: `breaks=c(30,50,70,100,150,200)`

(c) Suppose the top 15 households (who earned $150,000 to $200,000) each earn $4000 more. For each of the summary statistics below, say what happens to the summary statistic. Does it go up, go down, or stay the same? If possible, say how much it goes up or down by.

average: go up by $\frac{4000}{20}=200$ for this version or $\frac{2000}{20}=100$ for the other one

median: stay same

SD: go up

4. *(25 points:5 each)*
For admission into a selective elementary school, 500 kids were given a math and reading test. Assume both scores have histograms that follow the normal curve. The summary statistics are as follows:

	Average	SD
Math	76	4
Reading	82	5

The data is loaded into R as a data frame called `data` with two variables. The first variable is called `math` and has the math scores. The second variable is called `reading` and has the reading scores. You can assume that the data is already loaded in, and the dplyr package is attached (so you can use any functions in those packages).

(a) Use the normal curve to approximate the score at the 60th percentile on the math test (actually get an answer, don't write code).

$z \approx .25$

Score = 77

Score = 86

$E(x) + z_{score} * SD = 76 + (.25)*4$

$82 + 0.8(5) = 86$ $= 77$

(b) Write code in R to find the proportion of math scores that are above 70 on the math test.

mean (data$math > 70) or mean (select(data, math) > 70)

mean (data$math > 75)

(c) Write code in R to find the average reading score for the students that got over 70 on the math test.

mean(data $reading [data $ math > 70])

or

mean(filter(data, math > 70) $reading) mean(data$math [data $reading > 75])

(d) Suppose we choose 49 math scores out of the 500, with replacement. What is the chance that the average of these scores is 75 or more?

$SE_{avg} = \frac{SD_x}{\sqrt{n}} = \frac{4}{\sqrt{49}} = \frac{4}{7}$ $\frac{75-74}{\frac{4}{7}} = \frac{7}{4}$

z-value = 1.75

≈ 92% between -1.75 & 1.75

≈96%

$SE_{avg} = \frac{4}{6}$ $z = \frac{75-76}{4/3} = -3/2$ ≈93.3%

(e) Again suppose we choose 49 math scores out of the 500, with replacement. What is the chance that the percent of scores above the 60th percentile is 62% or more?

Assuming a) $SD = \sqrt{.6 \times .4} = \sqrt{.24} = .49$ $SD = \sqrt{(.21)(.79)} = .41$

$SE_{\%} = .0679$

$SE_x = \frac{.49}{\sqrt{49}} = \frac{.49}{7} = .7$ $\frac{.62-.60}{.07} = \frac{.02}{.07} = 2/7$

$\frac{.77-.79}{.0679} = -0.3$

$\Phi(-.3) \approx 62\%$

$\approx \Phi\left(\frac{2}{7}\right) \approx 38\text{-}39\%$