Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

September 28, 2023
52 total points

Problem Set 4 Solutions

1. (**6 points—3 each**) On the midterm exam in introductory statistics, an instructor always gives a grade of B to students who score between 80 and 90. The scores tend to have a normal distribution with a mean $\mu = 83$ and a standard deviation $\sigma = 5$. About what fraction of the students get a B?

   (a) First, answer this question using what you know about the normal distribution.

   (b) Now use simulated data in Stata. Generate 1,000 student exam scores—this instructor has a big class!—from a normal distribution with the above parameters. Then answer the question based on the data you drew. Are there any differences between your two answers?

   This question is asking: $P(80 \leq X \leq 90) = P(\frac{80-83}{5} \leq \frac{X-\mu}{\sigma} \leq \frac{90-83}{5}) = P(-0.6 \leq z \leq 1.4)$. For part (a), we can use Stata:

   `display normal(1.4)-normal(-0.6)`, this probability is 0.645. Or, about 64.5% of students get a B. See the attached do-file for part (b). The proportion scoring between 80 and 90 (inclusive) will generally differ from 0.645 because this is a random sample of 1,000.

2. (**8 points—2 each**) Suppose the SAT math scores of high school seniors follow a normal distribution with mean $\mu = 550$ and standard deviation $\sigma = 100$. Now suppose you take a sample of 25 seniors and calculate the sample mean ($\bar{x}$).

   (a) What is $E(\bar{x})$? What is $Var(\bar{x})$? What is the *standard error* of $\bar{x}$?

   $$\begin{aligned} E(\bar{x}) &= \mu = 550 \\ Var(\bar{x}) &= \sigma^2/n = 100^2/25 = 400 \\ se(\bar{x}) &= \sigma/\sqrt{n} = 100/\sqrt{25} = 20 \end{aligned}$$

   (b) Suppose you would like to increase precision and cut your standard error in half. What sample size would you need to use?

Currently the standard error of $\bar{x}$ is 20 with a sample size of 25. If we would like the standard error cut in half to 10, the sample size n will need to satisfy: $100/\sqrt{n} = 10$ or n=100. In general, with a standard deviation of σ and a desired standard error of se, the required sample size will be: $n = \sigma^2/se^2$.

(c) Based on the original sample size of 25, what is the probability that you draw a random sample with a $\bar{x}$ of 590 or higher?

Here we can apply our knowledge that SAT scores follow a normal distribution. The Central Limit Theorem says that if SAT scores are normal, then $\bar{x}$ drawn from a random sample of SAT scores will have a normal sampling distribution, with $E(\bar{x}) = \mu$ and $se(\bar{x}) = \sigma/\sqrt{n}$. So:

$$\begin{aligned} P(\bar{x} \geq 590) &= P\left(\frac{\bar{x} - 550}{100/\sqrt{25}} \geq \frac{590 - 550}{100/\sqrt{25}}\right) \\ &= P(z \geq 2) \\ &= 0.023 \end{aligned}$$

Or 2.3%. The last calculation comes from a statistical table or Stata (`display 1-normal(2)`)

(d) Based on the original sample size of 25, what is the probability what you draw a random sample with a $\bar{x}$ between 525 and 575?

By the same logic as part c:

$$\begin{aligned} P(525 \leq \bar{x} \leq 575) &= P\left(\frac{525 - 550}{100/\sqrt{25}} \leq \frac{\bar{x} - 550}{100/\sqrt{25}} \leq \frac{575 - 550}{100/\sqrt{25}}\right) \\ &= P(-1.25 \leq z \leq 1.25) \\ &= 0.789 \end{aligned}$$

Or 78.9%. The last calculation comes from a statistical table or Stata (`display normal(1.25)-normal(-1.25)`)

3. (**4 points**) A national survey conducted in June 2021 by the University of Southern California as part of its Understanding Coronavirus in America program asked participants whether they had received at least one dose of the coronavirus vaccine. Of 1,626 adults interviewed, 67.58% said *yes*. Find the estimated standard error for the sample proportion ($\hat{\pi}$) reporting they had received at least one dose of the vaccine. Interpret this in words. Hint: use the sample proportion in place of the *population* proportion (π) where required.

Recall the standard error of a sample proportion is $\sqrt{(\pi(1-\pi))/n}$. Since π is unknown we can substitute in our best estimate of π, which is $\hat{\pi}$. In this case the standard error is $\sqrt{(0.6758 * 0.3242)/1626} = 0.0116$. In any given random sample of 1,626 persons, the proportion who answer "yes" to this question will vary. The standard error 0.0116 is a measure of how much the proportion answering "yes" varies from sample to sample.

4. (**5 points**) Mr. Grumpy and Mr. Happy are both running for Governor. Mr. Grumpy will eventually win the election with 51 percent of the vote. A day before the election, a state-wide newspaper surveys 100 people about their choice for governor. Assume the survey respondents accurately report who they will vote for. What is the probability *Mr. Happy* will be supported by 51 percent or more of the survey respondents?

Knowing that Mr. Grumpy will eventually win with 51 percent of the vote means $\pi = 0.51$—the true population proportion of voters who favor Grumpy is 0.51. A random sample of 100 voters will provide the *sample* proportion $\hat{\pi}$, an estimate of the true population proportion. The standard error of a sample proportion across repeated samples is $\sqrt{(\pi(1-\pi))/n}$, or in this case, $\sqrt{(0.51 * 0.49)/100} = 0.050$. (Note this uses the known population proportion π since we know the ultimate result here is $\pi = 0.51$). The probability of drawing a random sample in which the other candidate (Happy) comes out on top with 51 percent or more of the survey respondents favoring that candidate is the same as the probability that *49 percent or fewer support Grumpy*. Applying the CLT, the sample proportion will have an approximate normal distribution with a large enough n. So:

$$\begin{aligned} P(\hat{\pi} \leq 0.49) &= P\left(\frac{\hat{\pi} - 0.51}{0.050} \leq \frac{0.49 - 0.51}{0.050}\right) \\ &= P(z \leq -0.4) \\ &= 0.345 \end{aligned}$$

In other words, there is a 34.5% chance that a random sample of 100 would show 49% or less support for Grumpy if the underlying population proportion was 51%. One could also solve this by defining $\pi = 0.49$ as the population proportion supporting Happy, and then finding $P(\hat{\pi} \geq 0.51)$. The result would be the same.

5. (**18 points**) The Chi-squared (χ^2) distribution is a common probability distribution in statistics defined by one parameter k (the degrees of freedom). If $x \sim \chi^2(k)$, then $E(x) = k$ and $Var(x) = 2k$. The skewness of the distribution is $\sqrt{8/k}$. For large values of k, the distribution is symmetric. For smaller values of k, it is positively skewed. For this problem, let $k = 10$. Create a Stata do file that does the following:

   See Stata results attached below.

   (a) (**2 points**) Draw 200 random values from this Chi-squared distribution. Note Stata has a random number function for the Chi-squared distribution.

   (b) (**2 points**) Create a histogram for your sample data and find the sample mean $\bar{x}$ and standard deviation s. How do these compare to the (known) population mean and standard deviation? Describe the shape of your distribution: is it symmetric or skewed? What is the skewness statistic?

   (c) (**2 points**) Following one of the methods shown in class, conduct a simulation that repeatedly samples 10 observations from the Chi-squared distribution (with $k = 10$ as before), a total of 100 times. Your simulation should store the sample mean $\bar{x}$ on each iteration. When complete, use your data to answer the next questions.

   (d) (**4 points**) Create a histogram for these simulated sample means. What is the mean of these values? What is the standard deviation? How do these compare to what you would have predicted them to be, before you drew any samples? Explain.

   (e) (**2 points**) Based on your simulated sampling distribution, what is the probability of drawing a sample with a $\bar{x}$ greater than 11?

   (f) (**6 points**) Repeat parts (c)-(e) but increase the random sample size to 50. How does this change your results?

6. (**5 points**) Consider this probability distribution for student absences from the last problem set:

# of Days	0	1	2	3	4	5	6	7	8
Probability	0.2	0.14	0.25	0.11	0.1	0.09	0.05	0.03	0.03

Now imagine you were to draw random samples of 50 students repeatedly from this population. What would the sampling distribution of $\bar{x}$ look like? What would be its mean? Its standard error? How do you know?

This population is definitely not a normal distribution. However, the CLT tells us that if the sample size is large enough, $\bar{x}$ will have an approximately normal sampling distribution with a mean of μ and a standard error of $\sigma/\sqrt{n}$. From the last problem set, $\mu = E(X) = 2.57$ and $\sigma = \sqrt{Var(X)} = \sqrt{4.57} = 2.14$ so that the standard error would be $2.14/\sqrt{50} = 0.302$.

7. (**6 points—2 each**) In Stata, use the code below to create a "population" distribution that reflects the probability distribution in part (6):

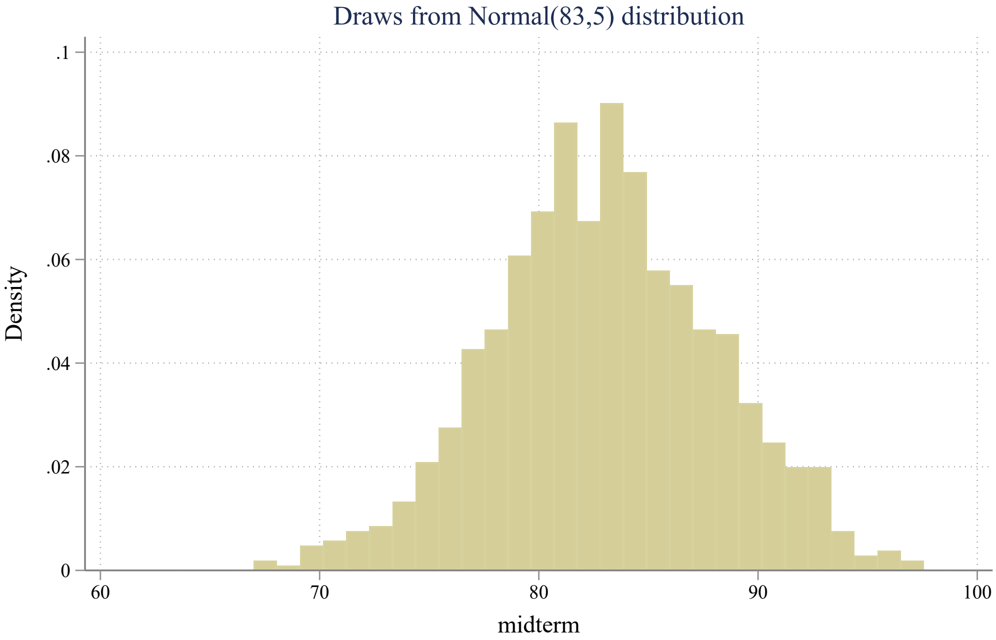
See Stata results attached below.

(a) What is the mean, median, variance, and standard deviation of this population?

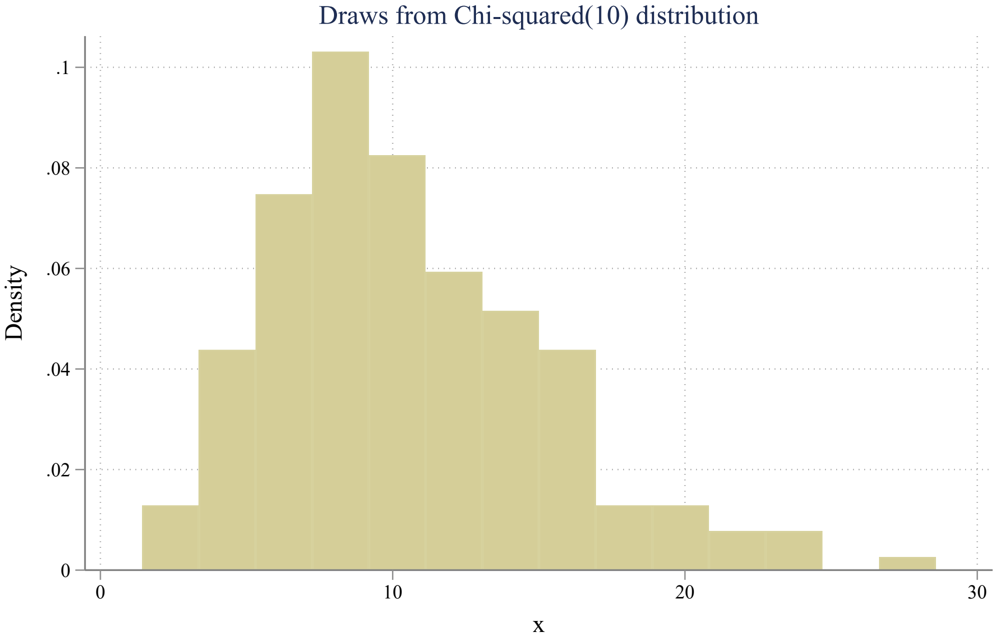
(b) Use the `bootstrap` prefix to draw 1,000 repeated samples of size $n = 30$ from this population, each time retaining the mean, median, and standard deviation. (Hint: you will `bootstrap` the `summarize, detail` command). Save the results in a new dataset.

(c) Provide histograms for your resulting means, medians, and standard deviations. Across your 1,000 samples, what the *average* mean, median, and standard deviation? For this simulation, what is the *standard error* of the mean, median, and standard deviation? Explain in words what this quantity represents.

```
.
. // *********************************************************************
. // LPO.8800 Problem Set 4 - Solutions to Questions 1, 5 and 7
. // Last updated: September 24, 2023
. // *********************************************************************
.
.         cd "$pset"
C:\Users\corcorsp\Dropbox\_TEACHING\Statistics I - PhD\Problem sets\Problem set 4
.
. // *************
. // Question 1
. // *************
.
.    // In population what proportion fall between 80 and 90? 0.645
.    display normal((90-83)/5) - normal((80-83)/5)
.64499022
.
.    // Now generate own data using 1,000 random draws from this population
.    set seed 1005
.    set obs 1000
number of observations (_N) was 0, now 1,000
.    gen midterm = rnormal(83, 5)
.    histogram midterm, title(Draws from Normal(83,5) distribution) ///      name(midterm,
> replace)
(bin=29, start=67.000526, width=1.0537246)
.    graph export midterm.pdf, as(pdf) replace
(file midterm.pdf written in PDF format)
```

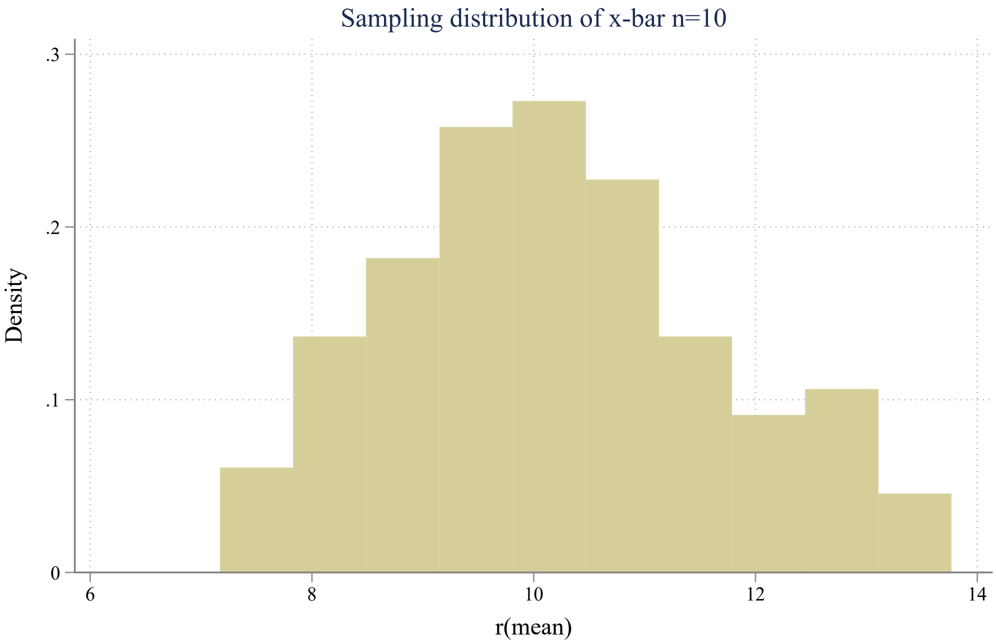


```
.    gen gradeb = midterm>=80 & midterm<=90
.    summ gradeb
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      gradeb |      1,000        .629    .4833142          0          1
.
. // In my case, 62.9% of scores fall between 80 and 90 (inclusive). (Your
. // results will vary due to the random draws). This differs from the known
. // proportion in the population that fall between 80-90, found above (64.5%).
. // They differ because the generated data represent one, albeit large, random
. // sample from the population.
.
.
. // *************
. // Question 5
. // *************
.
. // ********
. // Part a-b
. // ********
.
.         set seed 5001
.         clear
.         set obs 200
number of observations (_N) was 0, now 200
.         gen x=rchi2(10)
.         histogram x, title(Draws from Chi-squared(10) distribution) name(chi2x, replace)
(bin=14, start=1.4292891, width=1.9391582)
.         graph export chi2x.pdf, as(pdf) replace
(file chi2x.pdf written in PDF format)
```

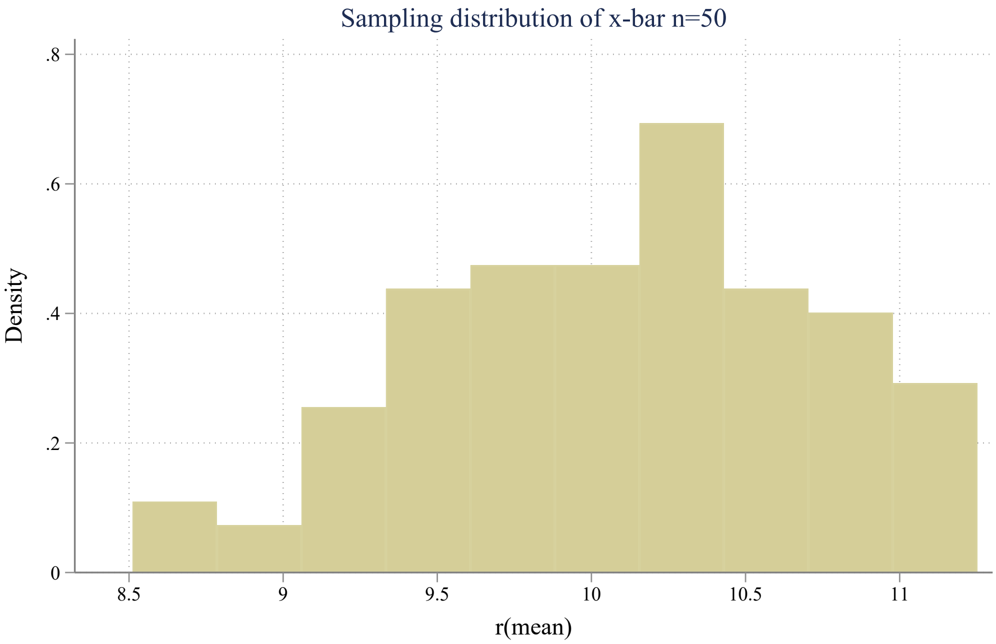


```
.         sum x,detail

                              x
-------------------------------------------------------------
      Percentiles      Smallest
 1%      2.226934       1.429289
 5%      3.870594        2.22456
10%      5.071997       2.229308       Obs                  200
25%      7.195452       2.817673       Sum of Wgt.          200

50%      9.874336                      Mean            10.55522
                        Largest       Std. Dev.       4.753164
75%      13.54217       24.09533
90%      16.45877       24.25708       Variance        22.59257
95%      19.22234       24.37455       Skewness        .8102286
99%      24.31581        28.5775       Kurtosis        3.811873
.
```

```
. // In my case, the sample mean and variance are 10.555 and 22.593, respectively.
. // (Your results will vary due to the random draws). The known population mean
. // of this distribution is E(x)=k = 10, which is close to my sample mean.
. // The known population variance is Var(x)=2k = 20, which is again close to
. // my sample variance. The distribution is right-skewed.
. // See the descriptive statistics and histogram.
.
. // ********
. // Part c
. // ********
. // Simulate 100 samples of n=10 from Chi-squared(10) distribution
.
.         capture program drop chi2
.         program chi2, rclass
  1.                 drop _all
  2.                 set obs 10
  3.                 gen x=rchi2(10)
  4.                 summ x
  5.                 return scalar mean=r(mean)
  6.         end
.
.         simulate mean=r(mean) , reps(100) nodots: chi2
      command:  chi2
         mean:  r(mean)
.
. // ********
. // Part d
. // ********
. // View sampling distribution (100 samples)
.
.         histogram mean, title(Sampling distribution of x-bar n=10) ///                na
> me(xbar10,replace)
(bin=10, start=7.1717663, width=.65956135)
.         graph export xbar10.pdf, as(pdf) replace
(file xbar10.pdf written in PDF format)
```



```
.         summ mean, det
                           r(mean)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     7.215956       7.171766
 5%     7.855759       7.260146
10%     8.320474       7.688992       Obs                 100
25%     9.157813       7.764164       Sum of Wgt.         100
50%      10.1359                      Mean           10.21342
                        Largest       Std. Dev.      1.501115
75%      11.1312       12.91886
90%     12.45014       13.24501       Variance       2.253347
95%     12.87675       13.43837       Skewness       .2824596
99%     13.60288       13.76738       Kurtosis        2.48551
.
. // In my case the mean of the sample means is 10.21 and the standard deviation
. // is 1.50. The Central Limit Theorem tells us that the mean of sampling
. // distribution of x-bar will be mu, or 10 in this case. The true (population)
. // standard deviation of the sampling distribution of x-bar (the standard
. // error) will be sigma/sqrt(n), or sqrt(20)/sqrt(10)= 1.41 in this case.
.
. // ********
. // Part e
. // ********
. // What fraction of sample means exceed 11?
.
.         count if mean>11
  29
.         display ‘r(N)’/_N
.29
.
. // We can look directly at the resulting sampling distribution to find the
. // proportion of times x-bar is greater than 11. In my case this is 29%.
.
.
. // ********
. // Part f
. // ********
. // Repeat parts c-e with sample size of 50
.
.         capture program drop chi2
.         program chi2, rclass
  1.                 drop _all
  2.                 set obs 50
  3.                 gen x=rchi2(10)
  4.                 summ x
  5.                 return scalar mean=r(mean)
  6.         end
.
.         simulate mean=r(mean) , reps(100) nodots: chi2
      command:  chi2
         mean:  r(mean)
.
.         histogram mean, title(Sampling distribution of x-bar n=50) name(xbar50,replace)
(bin=10, start=8.5110826, width=.27406311)
.         graph export xbar50.pdf, as(pdf) replace
(file xbar50.pdf written in PDF format)
```

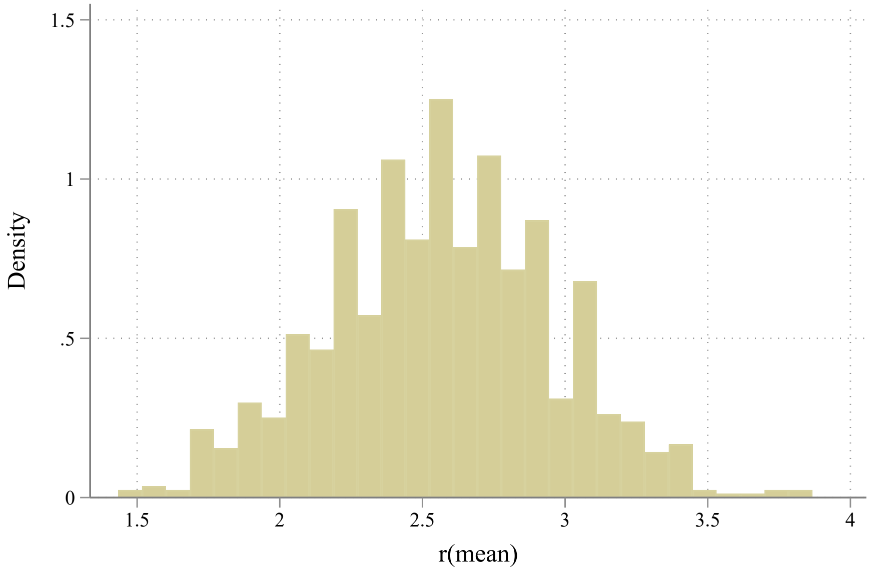


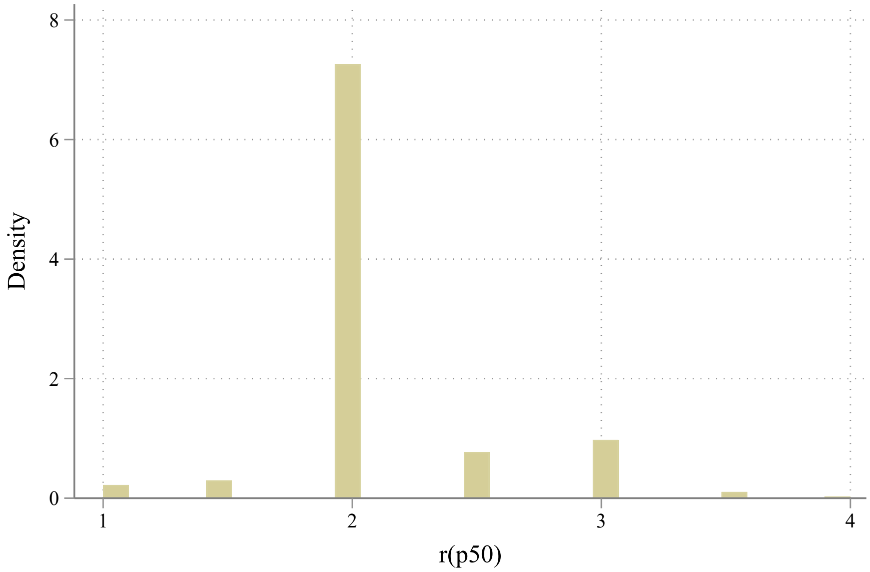
```
.         summ mean,det
                           r(mean)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     8.543072       8.511083
 5%     9.061373       8.575061
10%     9.227685       8.753469       Obs                 100
25%     9.653508       9.002769       Sum of Wgt.         100
50%     10.15116                      Mean           10.08912
                        Largest       Std. Dev.      .6313759
75%     10.54641       11.12409
90%     10.92034       11.14734       Variance       .3986355
95%     11.09847       11.19563       Skewness      -.2275299
99%     11.22367       11.25171       Kurtosis       2.444882
.
.         count if mean>11
  8
.         display ‘r(N)’/_N
.08
.
. // The result of a larger sample size is a sampling distribution with a smaller
. // standard error (now 0.63 instead of 1.5). The theoretical standard error
. // for a sample size of 50 is sqrt(20)/sqrt(50) = 0.632. In this new setting
. // the probability of drawing an x-bar>0.60 has fallen to 0.08 (i.e. such an
. // extreme value is less likely).
.
.
. // *************
. // Question 7
. // *************
. // Create a population distribution that looks like the probability distribution
. // in #6 (assuming 100 in the population for simplicity).
.
.         clear all
.         set obs 20
number of observations (_N) was 0, now 20
```

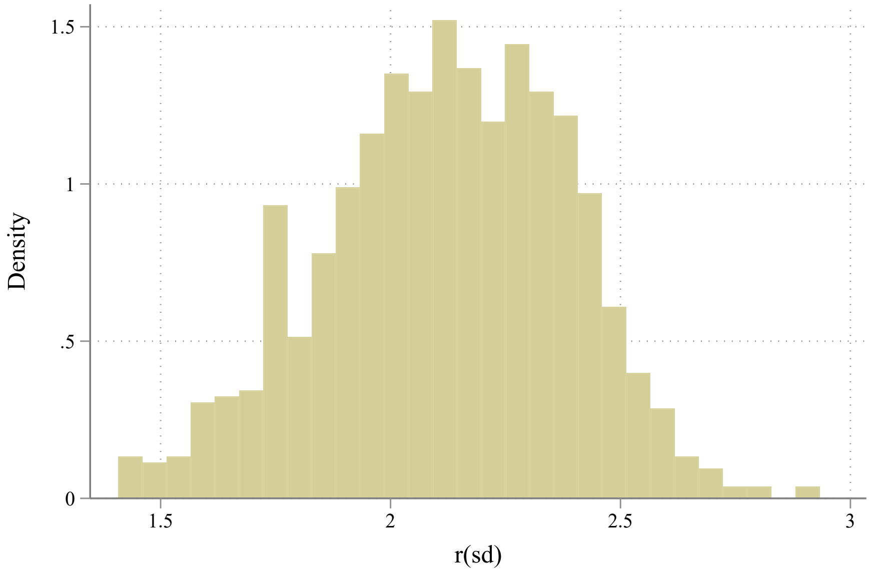
```
.         gen abs = 0
.         insobs 14 /* insobs inserts additional observations */
(14 observations added)
.         replace abs = 1 if abs==.
(14 real changes made)
.         insobs 25
(25 observations added)
.         replace abs = 2 if abs==.
(25 real changes made)
.         insobs 11
(11 observations added)
.         replace abs = 3 if abs==.
(11 real changes made)
.         insobs 10
(10 observations added)
.         replace abs = 4 if abs==.
(10 real changes made)
.         insobs 9
(9 observations added)
.         replace abs = 5 if abs==.
(9 real changes made)
.         insobs 5
(5 observations added)
.         replace abs = 6 if abs==.
(5 real changes made)
.         insobs 3
(3 observations added)
.         replace abs = 7 if abs==.
(3 real changes made)
.         insobs 3
(3 observations added)
.         replace abs = 8 if abs==.
(3 real changes made)
.         tabulate abs
        abs |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         20       20.00       20.00
          1 |         14       14.00       34.00
          2 |         25       25.00       59.00
          3 |         11       11.00       70.00
          4 |         10       10.00       80.00
          5 |          9        9.00       89.00
          6 |          5        5.00       94.00
          7 |          3        3.00       97.00
          8 |          3        3.00      100.00
------------+-----------------------------------
      Total |        100      100.00
.
.
. // ********
. // Part a
. // ********
.
```

```
.         summ abs, detail
                             abs
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 100
25%            1              0       Sum of Wgt.         100
50%            2                      Mean               2.57
                        Largest       Std. Dev.      2.147373
75%            4              7
90%            6              8       Variance       4.611212
95%            7              8       Skewness       .7216121
99%            8              8       Kurtosis       2.769535
.
. // The mean is 2.57, the median is 2, the variance is 4.61 and the standard
. // deviation is 2.15. If these data represent the population, then these are
. // the population parameters.
.
. // ********
. // Part b
. // ********
.
.         bootstrap r(mean) r(p50) r(sd), size(30) reps(1000) ///                saving(ab
> sresults,replace): summ abs,detail
(running summarize on estimation sample)
Warning:  Because summarize is not an estimation command or does not set
          e(sample), bootstrap has no way to determine which observations are
          used in calculating the statistics and so assumes that all
          observations are used.  This means that no observations will be
          excluded from the resampling because of missing values or other
          reasons.
          If the assumption is not true, press Break, save the data, and drop
          the observations that are to be excluded.  Be sure that the dataset
          in memory contains only the relevant data.
(note: file absresults.dta not found)
Bootstrap replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
```

```
..................................................  1000
Bootstrap results                               Number of obs     =        100
                                                Replications      =      1,000
      command:  summarize abs, detail
        _bs_1:  r(mean)
        _bs_2:  r(p50)
        _bs_3:  r(sd)
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |       2.57   .3907242     6.58   0.000     1.804195    3.335805
       _bs_2 |          2   .4159904     4.81   0.000     1.184674    2.815326
       _bs_3 |   2.147373   .2642287     8.13   0.000     1.629495    2.665252
------------------------------------------------------------------------------
.
.         use absresults, clear
(bootstrap: summarize)
.         rename _bs_1 xbar
.         rename _bs_2 xmedian
.         rename _bs_3 xstddev
.
.         summ xbar xmedian xstddev
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        xbar |      1,000    2.565367    .3907242   1.433333   3.866667
     xmedian |      1,000       2.125    .4159904          1          4
     xstddev |      1,000    2.120229    .2642287   1.407696   2.932968
.
.         histogram xbar, name(xbarabs, replace) nodraw
(bin=29, start=1.4333333, width=.08390804)
.         histogram xmedian, name(xmedabs, replace) nodraw
(bin=29, start=1, width=.10344828)
.         histogram xstddev, name(xstddev, replace) nodraw
(bin=29, start=1.4076964, width=.05259556)
.         graph combine xbarabs xmedabs xstddev, col(1) xsize(3) ysize(6)
.         graph export q7b.pdf, as(pdf) replace
(file q7b.pdf written in PDF format)
```


Density
1.5
1
.5
0
1.5
2
2.5
3
3.5
4
r(mean)


Density
8
6
4
2
0
1
2
3
4
r(p50)


Density
1.5
1
.5
0
1.5
2
2.5
3
r(sd)

```
.
. // The histograms (sampling distributions) are shown. The average value
. // for the mean (xbar) was 2.57 with a standard deviation (std error) of 0.391.
. // The average value for the median was 2.13 with a standard deviation (std
. // error) of 0.416. The average value for the std deviation was 2.12 with a
. // std deviation (standard error) of 0.264. These are good examples of three
. // statistics that vary from sample to sample. These simluated sampling
. // distributions give us a sense of their average value across samples and how
. // much they vary from the average from sample to sample. (The std error
. // captures that).
.
.         capture log close
```