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

September 30, 2021
49 total points

Problem Set 5 Solutions

---

1. (**3 points**) According to *Current Population Reports*, the population distribution of number of years of education for self-employed individuals in the United States has a mean of $\mu = 13.6$ and a standard deviation of $\sigma = 3.0$. Find the mean and *standard error* of the sampling distribution of $\bar{x}$ for a random sample of:

   (a) 9 individuals: $\mathbf{E}(\bar{\mathbf{x}}) = \mu = \mathbf{13.6}$ and $\mathbf{se}(\bar{\mathbf{x}}) = \sigma/\sqrt{\mathbf{n}} = \mathbf{3}/\sqrt{\mathbf{9}} = \mathbf{1}$

   (b) 36 individuals: $\mathbf{E}(\bar{\mathbf{x}}) = \mu = \mathbf{13.6}$ and $\mathbf{se}(\bar{\mathbf{x}}) = \sigma/\sqrt{\mathbf{n}} = \mathbf{3}/\sqrt{\mathbf{36}} = \mathbf{0.5}$

   (c) 100 individuals: $\mathbf{E}(\bar{\mathbf{x}}) = \mu = \mathbf{13.6}$ and $\mathbf{se}(\bar{\mathbf{x}}) = \sigma/\sqrt{\mathbf{n}} = \mathbf{3}/\sqrt{\mathbf{100}} = \mathbf{0.3}$

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 $\sqrt{(0.6758 * 0.3542)/1626} = 0.0116$. 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, $\sqrt{(0.51 * 0.49)/100} = 0.050$. 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 PDF for the continuous exponential distribution is $f(x) = \lambda e^{-\lambda x}$ where λ is a constant > 0. It can easily be shown that $E(x) = 1/\lambda$ and $Var(x) = 1/\lambda^2$. For this problem, let $\lambda = 2$. Create a Stata do file that does the following:

   See Stata results attached below.

   (a) (**2 points**) Draw 200 random values from this exponential distribution. Stata has a random number function for the exponential distribution, though it asks for a "scale parameter." This scale parameter you should use is actually $1/\lambda$, not λ.

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

   (c) (**2 points**) Following one of the methods shown in class, conduct a simulation that repeatedly samples 10 observations from the exponential distribution (with $\lambda = 2$ 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 0.6?

   (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	3	4	5	6	7
Probability	0.08	0.24	0.41	0.20	0.07

   Now imagine you were to draw random samples of 30 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, $\mu = E(X) = 4.94$ and $\sigma = \sqrt{Var(X)} = \sqrt{1.04} = 1.02$ so that the standard error would

be $1.02/\sqrt{30} = 0.186$.

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

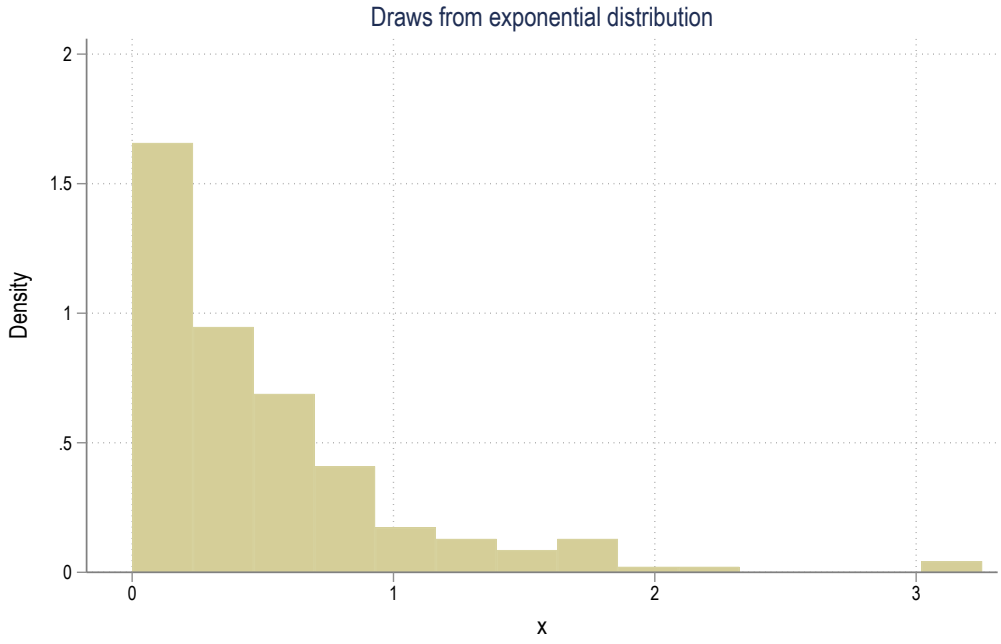
   See Stata results attached below.

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

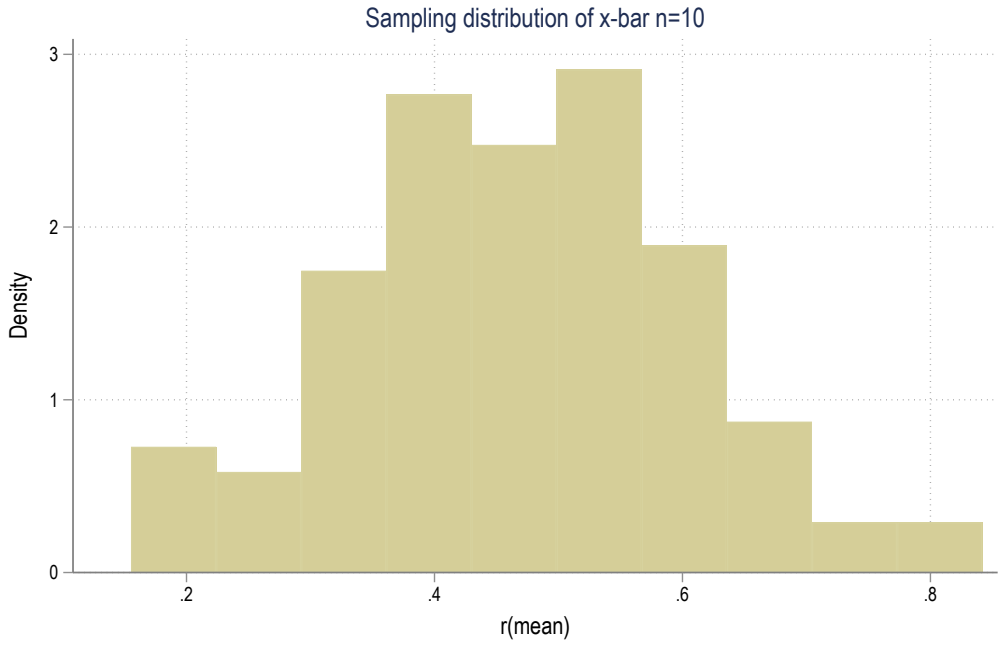
   (b) Use the `bootstrap` prefix to draw 1,000 repeated samples of size $n = 30$ from this population, each time retaining the mean, median, and standard deviation. (Hint: you will `bootstrap` the `summarize, detail` command). Save the results in a new dataset.

   (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 5 - Solutions to Questions 5 and 7
. // Last updated: September 25, 2021
. // ************************************************************************
.
. cd "$db\_TEACHING\Statistics I - PhD\Problem sets\Problem set 5"
C:\Users\corcorsp\Dropbox\_TEACHING\Statistics I - PhD\Problem sets\Problem set 5
.
. // *************
. // Question 5
. // *************
.
. // ********
. // Part a-b
. // ********
. set seed 5001
. clear
. set obs 200
number of observations (_N) was 0, now 200
. gen x=rexponential(0.5)
. histogram x, title(Draws from exponential distribution) name(expx, replace)
(bin=14, start=.00083047, width=.23229243)
. graph export expx.pdf, as(pdf) replace
(file expx.pdf written in PDF format)
```

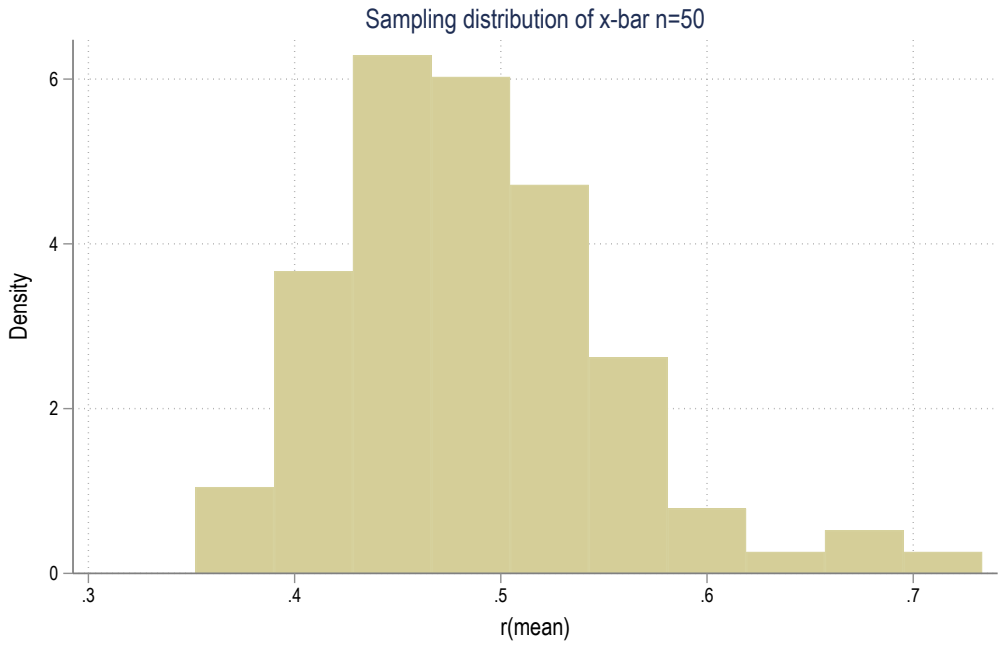


```
. sum x,detail
                            x
-------------------------------------------------------------
      Percentiles      Smallest
 1%      .003129       .0008305
 5%     .0233381       .0017492
10%     .0478808       .0045087       Obs                 200
25%     .1302637       .0048741       Sum of Wgt.         200
50%     .3865653                      Mean           .5003724
                        Largest       Std. Dev.      .5238797
75%     .6567973       1.883473
90%     1.122045       2.298275       Variance       .2744499
95%     1.622603       3.242462       Skewness       2.238356
99%     2.770368       3.252924       Kurtosis       10.08371
.
. // In my case, the sample mean and variance are 0.5004 and 0.2744, respectively.
. // (Your results will vary due to the random draws). The known population mean
. // of this distribution is E(x)=1/lambda = 1/2, which is very close to my sample
. // mean. The known population variance is Var(x)=1/lambda^2 = 1/4, which is
. // again close to my sample variance. The distribution is very right-skewed.
. // See the descriptive statistics and histogram.
.
. // ********
. // Part c
. // ********
. // Simulate 100 samples of n=10 from exponential distribution
.
. capture program drop exp
. program exp, rclass
  1.         drop _all
  2.         set obs 10
  3.         gen x=rexponential(0.5)
  4.         summ x
  5.         return scalar mean=r(mean)
  6. end
.
. set seed 4321
. simulate mean=r(mean) , reps(100) nodots: exp
      command:  exp
         mean:  r(mean)
.
. // ********
. // Part d
. // ********
.
. histogram mean, title(Sampling distribution of x-bar n=10) name(xbar10,replace)
(bin=10, start=.15519677, width=.06865803)
. graph export xbar10.pdf, as(pdf) replace
(file xbar10.pdf written in PDF format)
```




```
. summ mean, det
                          r(mean)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .1559954       .1551968
 5%     .2365946        .156794
10%     .3076856       .1827664       Obs                 100
25%     .3784107       .1983413       Sum of Wgt.         100
50%     .4726311                      Mean           .4702106
                        Largest       Std. Dev.       .139157
75%     .5584147       .7404677
90%     .6355464       .7421322       Variance       .0193647
95%      .694655       .8042927       Skewness       .0509755
99%     .8230349       .8417771       Kurtosis       2.946249
.
. // In my case the mean of the sample means is 0.470 and the standard deviation
. // is 0.139. The Central Limit Theorem tells us that the mean of sampling
. // distribution of x-bar will be mu, or 0.5 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(25)/sqrt(10)= 0.158 in this case.
.
. // ********
. // Part e
. // ********
.
. count if mean>0.6
  18
. display `r(N)'/_N
.18
.
. // We can look directly at the resulting sampling distribution to find the
. // proportion of times x-bar is greater than 0.6. In my case this is 18%.
.
.
. // ********
. // Part f
```

```
. // ********
. // Repeat parts c-e with sample size of 50
.
. capture program drop exp
. program exp, rclass
  1.         drop _all
  2.         set obs 50
  3.         gen x=rexponential(0.5)
  4.         summ x
  5.         return scalar mean=r(mean)
  6. end
.
. set seed 4321
. simulate mean=r(mean) , reps(100) nodots: exp
      command:  exp
         mean:  r(mean)
.
. histogram mean, title(Sampling distribution of x-bar n=50) name(xbar50,replace)
(bin=10, start=.3519119, width=.038156)
. graph export xbar50.pdf, as(pdf) replace
(file xbar50.pdf written in PDF format)
```



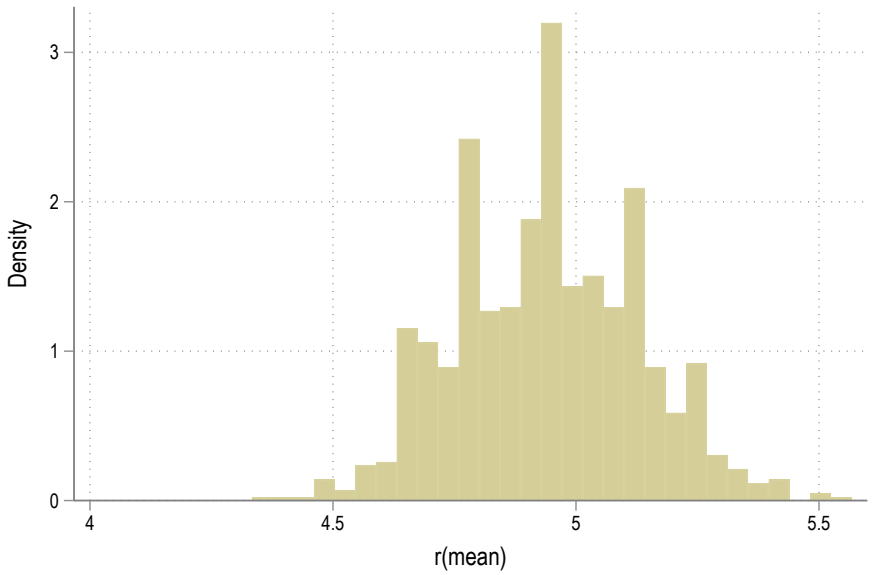

```
. summ mean,det
                          r(mean)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .3586884       .3519119
 5%     .3972536       .3654648
10%     .4061931       .3768631       Obs                 100
25%     .4430249       .3799605       Sum of Wgt.         100
50%     .4797811                      Mean           .4868304
                        Largest       Std. Dev.      .0675938
75%      .525467       .6217108
90%     .5751603       .6639152       Variance       .0045689
95%     .6015339         .66696       Skewness       .7897101
99%     .7002159       .7334719       Kurtosis       4.119267
```

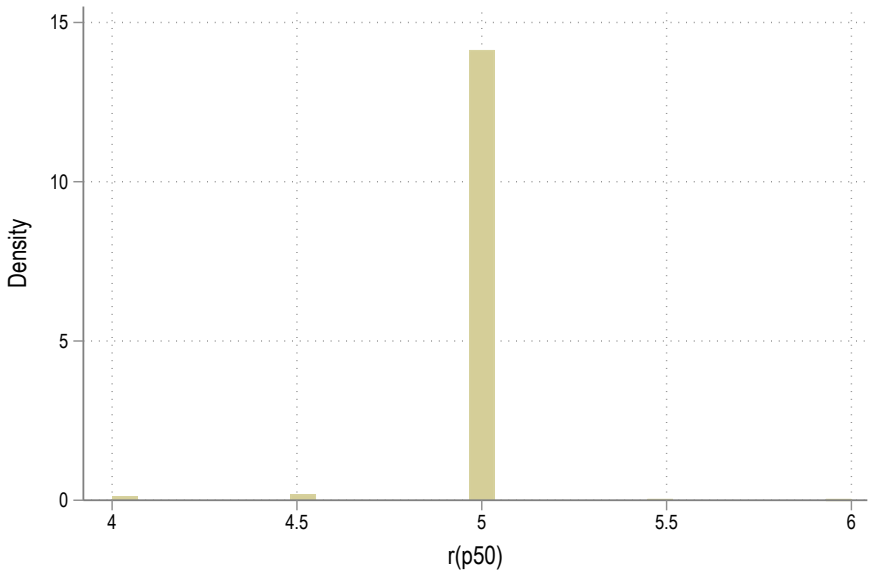
```
.
. count if mean>0.6
  5
. display ‘r(N)’/_N
.05
.
. // The result of a larger sample size is a sampling distribution with a smaller
. // standard error (now 0.068 instead of 0.139). The theoretical standard error
. // for a sample size of 50 is sqrt(0.25)/sqrt(50) = 0.071. In this new setting
. // the probability of drawing an x-bar>0.60 has fallen to 0.05 (i.e. such an
. // extreme value is less likely).
.
.
. // *************
. // Question 7
. // *************
.
. clear all
. set seed 5002
. set obs 8
number of observations (_N) was 0, now 8
. gen abs = 3
. insobs 24
(24 observations added)
. replace abs = 4 if abs==.
(24 real changes made)
. insobs 41
(41 observations added)
. replace abs = 5 if abs==.
(41 real changes made)
. insobs 20
(20 observations added)
. replace abs = 6 if abs==.
(20 real changes made)
. insobs 7
(7 observations added)
. replace abs = 7 if abs==.
(7 real changes made)
.
. // ********
. // Part a
. // ********
. summ abs, detail
                            abs
-------------------------------------------------------------
      Percentiles      Smallest
 1%             3              3
 5%             3              3
10%             4              3       Obs                 100
25%             4              3       Sum of Wgt.         100
50%             5                      Mean               4.94
                        Largest       Std. Dev.      1.023166
75%             6              7
90%             6              7       Variance       1.046869
95%             7              7       Skewness       .0632816
99%             7              7       Kurtosis       2.638078
.
. // The mean is 4.94, the median is 5, the variance is 1.047 and the standard
. // deviation is 1.023. If these data represent the population, then these are
. // the population parameters.
```

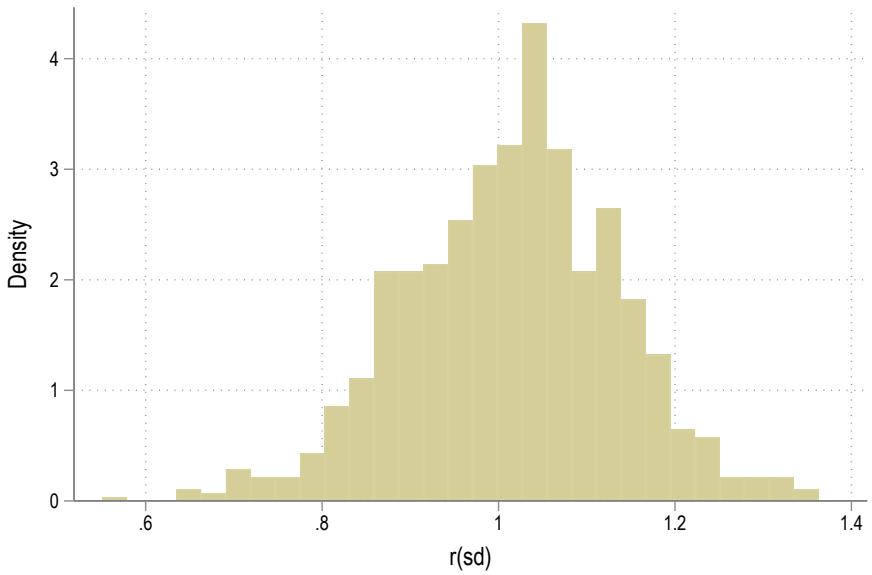
```
.
. // ********
. // Part b
. // ********
. bootstrap r(mean) r(p50) r(sd), size(30) reps(1000) saving(absresults,replace): summ abs
> ,detail
(running summarize on estimation sample)
Warning:  Because summarize is not an estimation command or does not set
          e(sample), bootstrap has no way 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.
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 |       4.94    .188164    26.25   0.000     4.571205    5.308795
       _bs_2 |          5    .116539    42.90   0.000     4.771588    5.228412
       _bs_3 |   1.023166   .1196835     8.55   0.000     .7885907    1.257741
------------------------------------------------------------------------------
.
. 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    4.940767     .188164   4.333333   5.566667
     xmedian |      1,000      4.9865     .116539          4          6
     xstddev |      1,000    1.014101    .1196835   .5508614   1.362891
.
. histogram xbar, name(xbarabs, replace) nodraw
(bin=29, start=4.3333335, width=.04252873)
. histogram xmedian, name(xmedabs, replace) nodraw
(bin=29, start=4, width=.06896552)
. histogram xstddev, name(xstddev, replace) nodraw
(bin=29, start=.55086142, width=.02800101)
. 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
0
1
2
3
4
4.5
5
5.5
r(mean)


Density
0
5
10
15
4
4.5
5
5.5
6
r(p50)


Density
0
1
2
3
4
.6
.8
1
1.2
1.4
r(sd)

```
.
. // The histograms (sampling distributions) are attached below. The average value
. // for the mean (xbar) was 4.94 with a standard deviation (std error) of 0.188.
. // The average value for the median was 4.987 with a standard deviation (std
. // error) of 0.117. The average value for the variance was 1.014 with a standard
. // deviation (standard error) of 0.1197. 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
```