Lab 4 - Sampling distributions

Paranormal Distribution

7/21/16

Lab report

Load data:

```
load(url("https://stat.duke.edu/~mc301/data/ames.RData"))
```

Set a seed:

```
set.seed(578498)
```

Exercises:

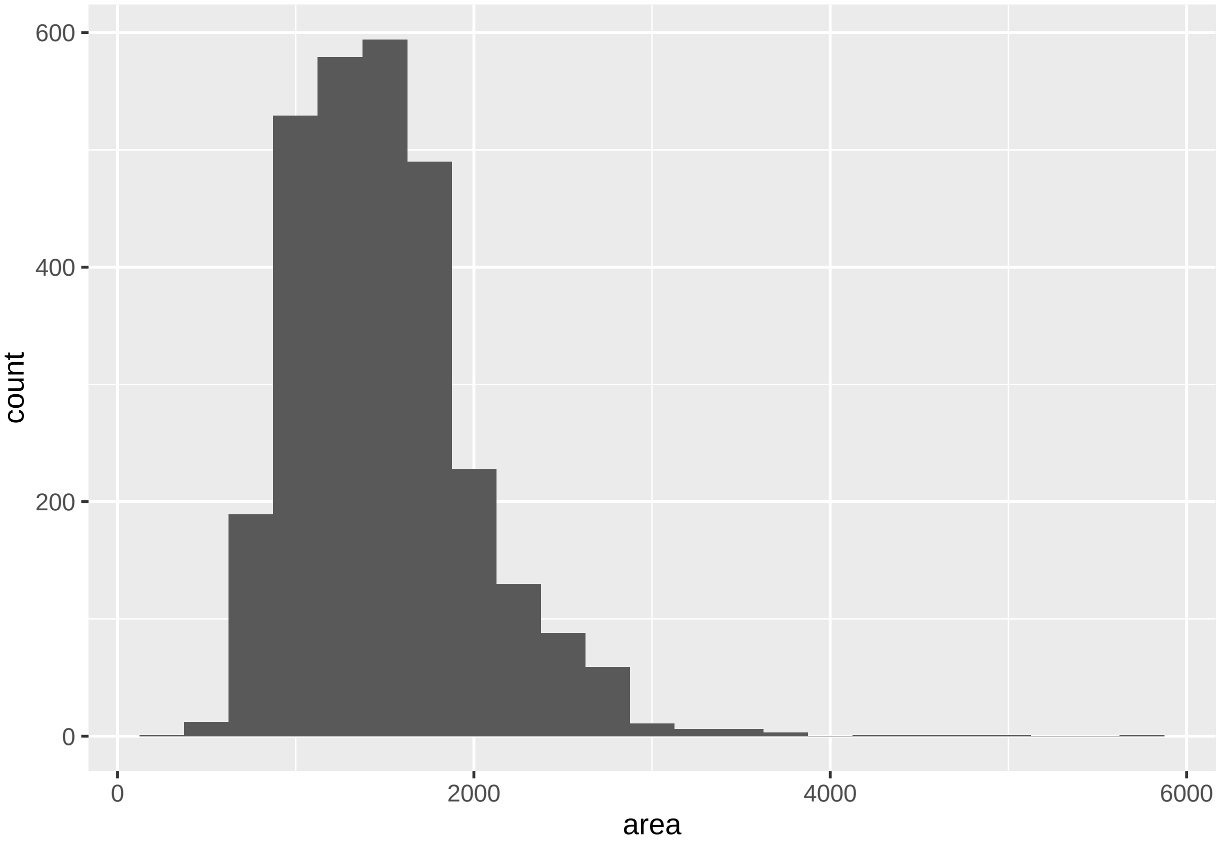
Exercise 1:

The distribution of area from the "ames" data frame is unimodal and right-skewed with no clear outliers. The mean is 1499.69, but, due to skewness, the median provides a more robust description of center. The median is 1442 and the IQR is 616.75. The right-tail shows considerably greater variance (3rd quartile = 1742.75, maximum = 5642) and the left-tail levels off around 500 (1st quartile = 1126. minimum = 334).

```
ames %>%
  summarise(mu = mean(area), pop_med = median(area),
            sigma = sd(area), pop_iqr = IQR(area),
            pop_min = min(area), pop_max = max(area),
            pop_q1 = quantile(area, 0.25),  # first quartile, 25th percentile
            pop_q3 = quantile(area, 0.75))  # third quartile, 75th percentile
```

```
## # A tibble: 1 x 8
##        mu pop_med    sigma pop_iqr pop_min pop_max pop_q1  pop_q3
##     <dbl>   <dbl>    <dbl>   <dbl>   <int>   <int>  <dbl>   <dbl>
## 1 1499.69    1442 505.5089  616.75     334    5642   1126 1742.75
```

```
qplot(data = ames, x = area, binwidth = 250, geom = "histogram")
```



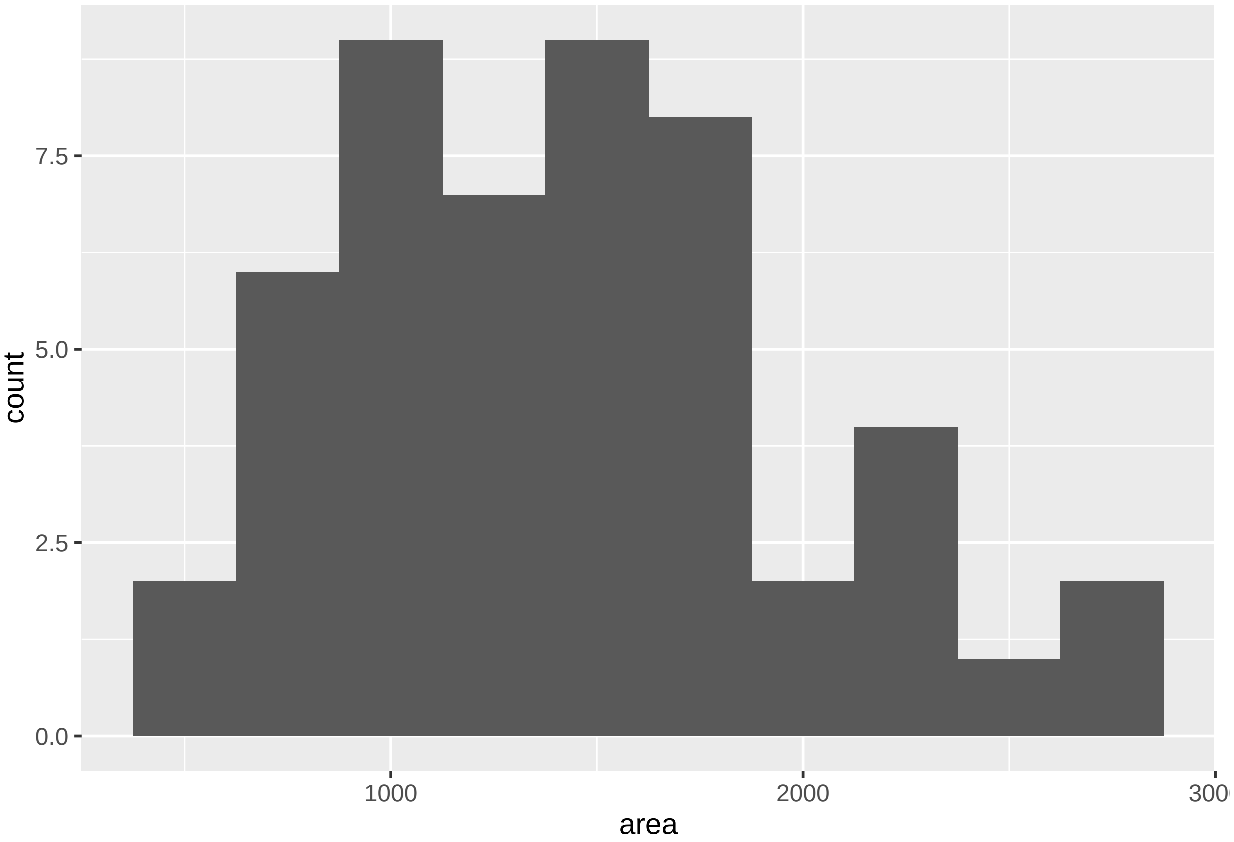
Exercise 2:

The distribution of area from "samp1" loosely mirrors the population distribution in modality and skew. The unimodal shape is less clear in the sample, and two modes can be seen at the 1000 and 1500 bins. The sample distribution also shows greater variability (sam_iqr = 729 v. pop_iqr = 616.75) and a slightly different center (sam_med = 1406 v. pop_med = 1442).

```r
set.seed(1039)
samp1 <- ames %>%
  sample_n(50)
samp1 %>%
  summarise(x = mean(area), sam_med = median(area),
            s = sd(area), sam_iqr = IQR(area),
            sam_min = min(area), sam_max = max(area),
            sam_q1 = quantile(area, 0.25),  # first quartile, 25th percentile
            sam_q3 = quantile(area, 0.75))  # third quartile, 75th percentile
```

```
## # A tibble: 1 x 8
##         x sam_med       s sam_iqr sam_min sam_max sam_q1 sam_q3
##     <dbl>   <dbl>   <dbl>   <dbl>   <int>   <int>  <dbl>  <dbl>
## 1 1445.64    1406 547.813     729     480    2790   1041   1770
```

```r
qplot(data = samp1, x = area, binwidth = 250, geom = "histogram")
```



Exercise 3:

Sample means taken by two different teams are unlikely to match because of expected sampling variety. The difference between the two sample means is unlikely to be drastic, however, since the samples are drawn from identical populations and are sufficiently large (50) to yield relatively consistent results.

Exercise 4:

The mean of samp2 is slightly different from that of samp1 (1446.8 v. 1445.64). The samp2 mean is closer to the ames population mean (mu = 1499.69).

A sample of 100 would provide a more accurate point estimate of the population mean due to it's larger size. The larger a sample, the less variable the data and the more representative the sample is of the population. For this reason, a sample of size 1000 would provide the most accurate estimate of the population mean.

```
set.seed(85711)
samp2 <- ames %>%
  sample_n(50)
samp2 %>%
  summarise(x = mean(area), sam_med = median(area),
            s = sd(area), sam_iqr = IQR(area),
            sam_min = min(area), sam_max = max(area),
            sam_q1 = quantile(area, 0.25),  # first quartile, 25th percentile
            sam_q3 = quantile(area, 0.75))  # third quartile, 75th percentile
```

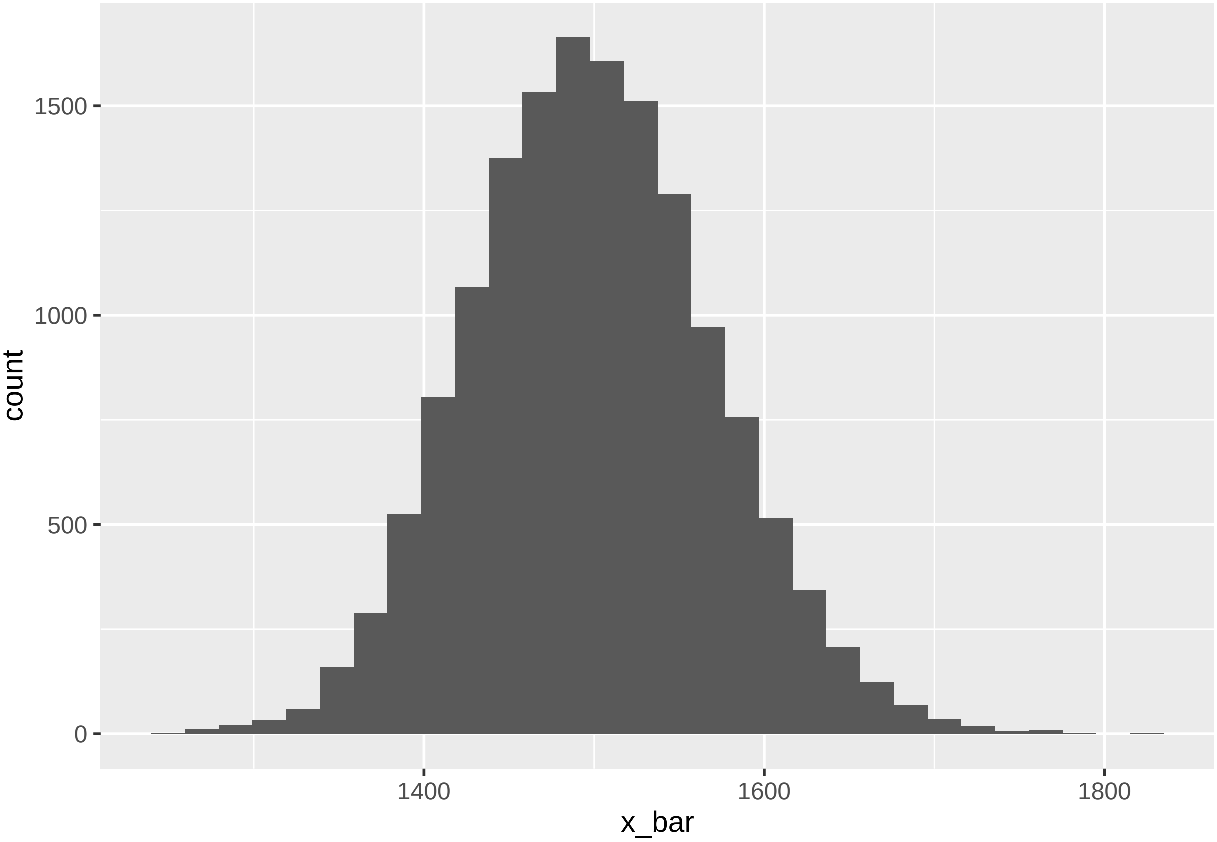
```
## # A tibble: 1 x 8
##         x sam_med        s sam_iqr sam_min sam_max  sam_q1  sam_q3
##     <dbl>   <dbl>    <dbl>   <dbl>   <int>   <int>   <dbl>   <dbl>
## 1 1446.8    1388 516.2487     722     438    3078 1057.75 1779.75
```

Exercise 5:

There are 15000 elements in sample_means50; each is a different sample of size 50 drawn from the ames population. The sampling distribution is nearly normal and centered at the mean of 1499.126. The standard deviation is 71.3.

```
set.seed(898989)
sample_means50 <- ames %>%
  rep_sample_n(size = 50, reps= 15000, replace = TRUE) %>%
  summarise(x_bar=mean(area))
qplot(data = sample_means50, x = x_bar)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Exercise 6:

There are 25 observations in sample_means_small. Each observation represents the mean of 10 randomly selected house areas drawn from the population with replacement.

```
set.seed(12345)
sample_means_small <- ames %>%
```

```r
  rep_sample_n(size = 10, reps = 25, replace = TRUE) %>%
  summarise(x_bar = mean(area))

print(sample_means_small)
```

```
## # A tibble: 25 x 2
##    replicate  x_bar
##        <int>  <dbl>
## 1          1 1539.9
## 2          2 1575.6
## 3          3 1503.3
## 4          4 1670.2
## 5          5 1438.5
## 6          6 1320.5
## 7          7 1588.8
## 8          8 1358.1
## 9          9 1446.0
## 10        10 1599.6
## # ... with 15 more rows
```

Exercise 7:

Each observation in the sampling distribution represents one mean from a random sample drawn from the population with replacement. As the sample size increases, standard error decreases, the distribution becomes more normal, and the sampling mean approaches the population mean (mu) of 1499.69. As the number of samples increases, the distribution follows a similar pattern: standard deviation decreases, the distribution grows more symmetric, and the sampling mean approaches 1499.69.

On your own:

1:

Based on the random sample of 15 below, the best point estimate of the population mean for ames prices is 199,901.7 (x_bar). Due to small sample size, this value may differ significantly from the true population mean for prices.

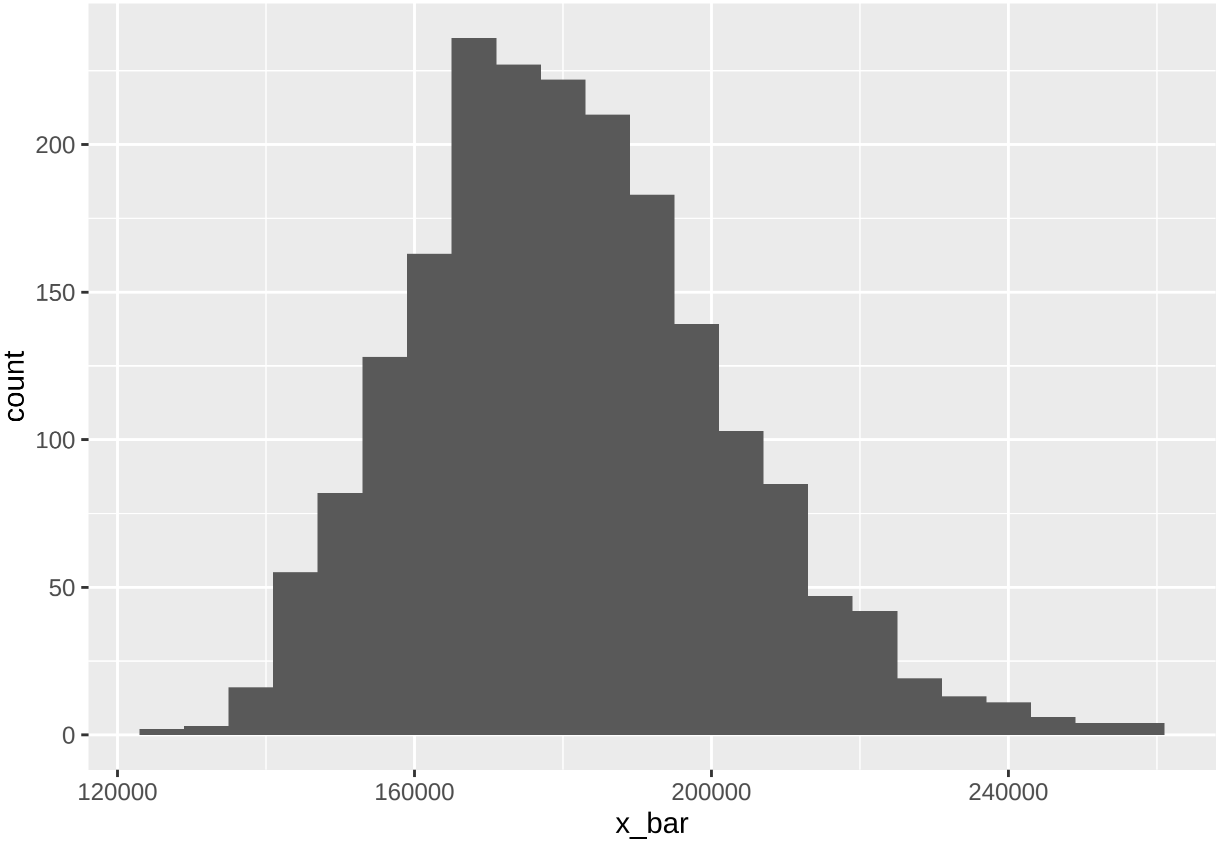
```r
set.seed(1234567)
samp1_price <- ames %>%
  sample_n(15) %>%
  summarise(x_bar = mean(price))
```

2:

The sampling distribution of "sample_means15" is unimodal and roughly symmetric with slightly greater variability on the right tail (slightly right-skewed). The center (mean) is 181,071.9 and the standard deviation is 20,330.6. Based on this distribution, we might estimate the population mean for ames prices to be 181,071.9 (the mean of the sampling distribution). This estimated mean differs from the true population mean (mu = 180,796.1).

```
set.seed(9865)
sample_means15 <- ames %>%
  rep_sample_n(size = 15, reps = 2000, replace = TRUE) %>%
  summarise(x_bar = mean(price))

qplot(data = sample_means15, binwidth = 6000, x = x_bar)
```



```
ames %>%
  summarise(mu = mean(price))
```

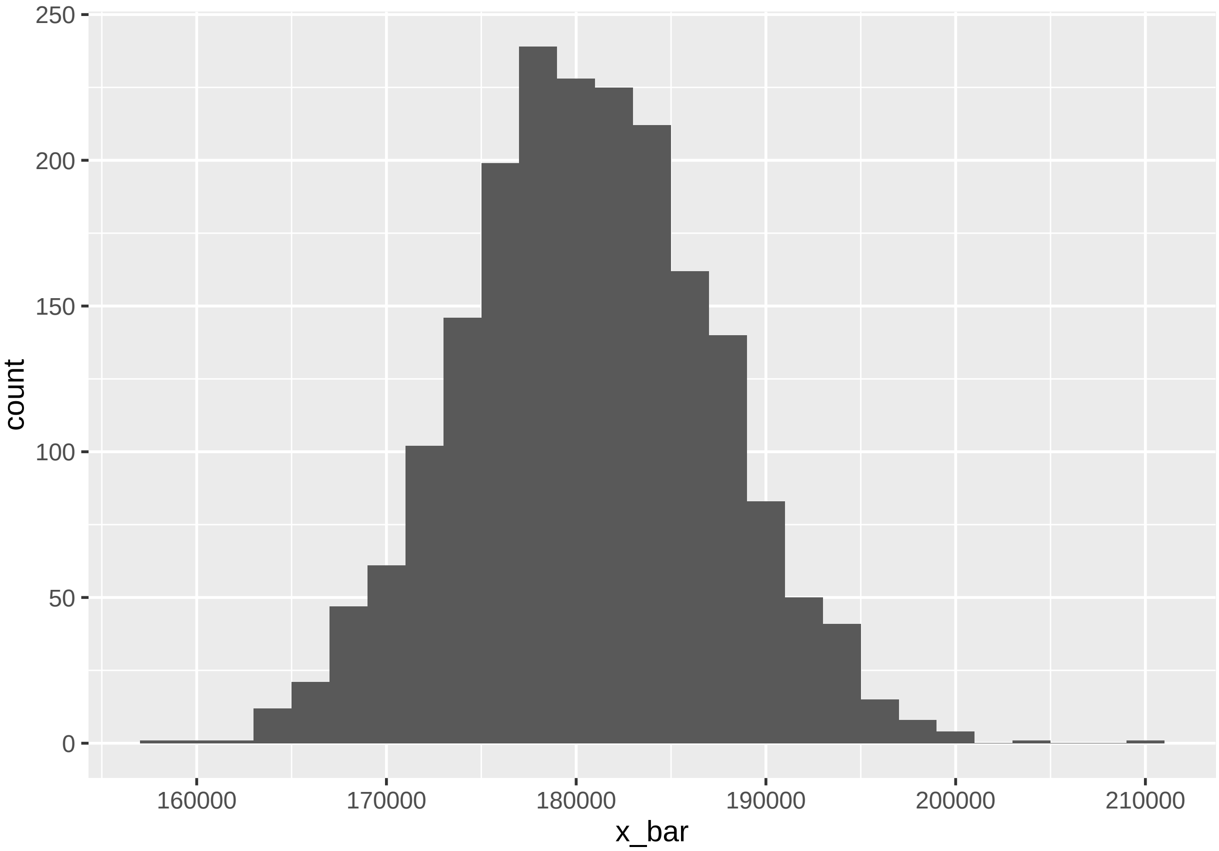
```
## # A tibble: 1 x 1
##         mu
##      <dbl>
## 1 180796.1
```

3:

The sampling distribution for "sample_means150" more closely resembles a normal curve than the distribution for "sample_means15." The larger sample size (n = 150) yields a significantly smaller standard deviation (6642.0 v. 21,053.0) and a more accurate sampling mean (180,606.0 v. 181,071.9). Based on the sampling distribution for "sample_means150," we might estimate the population mean for ames prices to be 180,606 (the mean of the sampling distribution).

```
set.seed(564565)
sample_means150 <- ames %>%
  rep_sample_n(size = 150, reps = 2000, replace = TRUE) %>%
  summarise(x_bar = mean(price))

qplot(data = sample_means150, binwidth = 2000, x = x_bar)
```



4:

The distribution of sample size 150 has a smaller spread. Because both distributions are roughly symmetric, this spread is best represented by standard deviation. The standard deviation for “sample_means150” is less than half that of “sample_means15” (6642.0 v. 21,053.0).

When trying to make accurate sampling estimates of population parameters, a sampling distribution with a smaller spread is preferable. Sampling distributions with small spreads yield more accurate, representative point estimates of center.

Teamwork report

Team member	Attendance	Author	Contribution %
Luul Lampkins	Yes	Yes	25%
Katie Payne	Yes	No	25%

Team member	Attendance	Author	Contribution %
Logan Laguna Kirkpatrick	Yes	Yes	25%
Weikuan Ji	Yes	No	25%
Total			100%