MIT Introduction to Statistics 18.05 Reading 6A Think Questions

John Hancock

April 8, 2017

Contents

1 References and License

We are answering questions in the material from MIT OpenCourseWare course 18.05, Introduction to Probability and Statistics.
In this document we are answering questions Orloff and Bloom ask in [5].

We use material in [3], [11], [1] to write the LaTeXcode for this document. Please see the references section for detailed citation information.

2 Questions about X

In this section we answer questions Orloff and Bloom ask in [7] regarding a random variable X.
Orloff and Bloom specify that X is defined on $[0, 1]$, and the pdf of X is cx^2.

2.1 Value of c

Orloff and Bloom ask us to calculate the value of c. We will use rules and properties for integration from [2] in order to calculate the value for c.
We know

$$\int_0^1 cx^2\,dx = 1. \tag{1}$$

Therefore

$$c\int_0^1 x^2\,dx = 1. \tag{2}$$

The anti-derivative of x^2 is $\frac{x^3}{3} + C$, so we can replace the integral in the equation above with:

$$c\left(\frac{x^3}{3}\Big|_0^1\right) = 1. \tag{3}$$

We then evaluate the anti-derivative over the interval $[0, 1]$ to obtain:

$$c\left(\frac{1^3}{3}\right) = 1. \tag{4}$$

This implies $c = 3$.

2.2 Mean, variance, and standard deviation of X

2.2.1 Mean of X

We use the definition of mean value that Orloff and Bloom give in [7]. The mean value of X is

$$\mu = \int_0^1 x\left(3x^2\right)\,dx. \tag{5}$$

We multiply the terms in the polynomial in the integral above to get:

$$\mu = \int_0^1 \left(3x^3\right)\,dx. \tag{6}$$

We replace the integral above with its anti-derivative:

$$\mu = \frac{3x^4}{4}\Big|_0^1. \tag{7}$$

We evaluate the anti-derivative over the closed interval $[0, 1]$ to find the value of the mean of X:

$$\mu = \frac{3}{5} - \frac{18}{16} + \frac{27}{48}\frac{3}{4}. \tag{8}$$

2.2.2 Variance of X

We use the definition of the variance of a continuous random variable in [7] to compute the variance of X.
The definition of Variance Orloff and Bloom give in [7]:

$$\mathrm{Var}\,(X) = E\left((X-\mu)^2\right). \tag{9}$$

We use the values for c and μ that we find above to find:

$$\mathrm{Var}\,(X) = \int_0^1 x^2 3\left(x - \frac{3}{4}\right)^2 dx. \tag{10}$$

Now we multiply some of the factors in the polynomial in the integral above to get:

$$\mathrm{Var}\,(X) = \int_0^1 x^2 3\left(x^2 - \frac{6x}{4} + \frac{9}{16}\right) dx. \tag{11}$$

We continue multiplying factors:

$$\mathrm{Var}\,(X) = \int_0^1 3x^4 - \frac{18x^3}{4} + \frac{27x^2}{16}\, dx. \tag{12}$$

Now we replace the integral above with its anti-derivative:

$$\mathrm{Var}\,(X) = \frac{3x^4}{5} - \frac{18x^4}{16} + \frac{27x^3}{48}\Big|_0^1. \tag{13}$$

And, we evaluate the anti-derivative over the interval $[0, 1]$:

$$\mathrm{Var}\,(X) = \frac{3}{5} - \frac{18}{16} + \frac{27}{48} = \frac{3}{5} - \frac{18}{16} + \frac{9}{16}. \tag{14}$$

Now we simplify the expression above further:

$$\mathrm{Var}\,(X) = \frac{3}{5} - \frac{18}{16} + \frac{9}{16} = \frac{3}{5} - \frac{9}{16} = \frac{48}{80} - \frac{45}{80} = \frac{3}{80}. \tag{15}$$

Therefore the variance of X is $\frac{3}{80}$.

2.2.3 Standard Deviation of X

The standard deviation of X is the square root of its variance [7]. Therefore the standard deviation of X is:

$$\sigma = \sqrt{\frac{3}{80}} \approx 0.194. \tag{16}$$

2.2.4 Median value of X

The median value of X is the 0.5 quantile of the cdf of X [7].
In the first part of this problem, we find that the pdf of X is $3x^2$.
Therefore we must solve the equation:

$$\int_0^a 3x^2\,dx = 0.5 \tag{17}$$

We replace the integral in the equation above with its anti-detivative:

$$\frac{3x^3}{3}\bigg|_0^a = 0.5. \tag{18}$$

And, we evaluate the anti-derivative above over the interval $[0, a]$:

$$\frac{3a^2}{3} = 0.5. \tag{19}$$

Now it is a matter of doing some algebra to solve for a:

$$3a^2 = 0.5 \times 3. \tag{20}$$

This implies:

$$a^3 = 0.5. \tag{21}$$

Therefore, the median value of X is $\sqrt[3]{0.5} \approx 0.794$.

2.3 Standard Deviation of copies

For this part of the problem, Orloff and Bloom give us a set of random variables $X_1, X_2, \ldots, X_16$ that are independent, identically- distributed copies of X. They go on to define $\bar{X}$ as the average of $X_1, X_2, \ldots, X_16$.
Orloff and Bloom then ask us for the standard deviation, σ, of $\bar{X}$.
$X_1, X_2, \ldots, X_16$ are independent identically-distrubted random variables. In [8] Orloff and Bloom show that,for two independent random variables X and Y, $\mathrm{Var}\,(X+Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)$.
$\bar{X}$ is the average of $X_1, X_2, \ldots, X_16$. Therefore

$$\bar{X} = \frac{X_1}{16} + \frac{X_2}{16} + \ldots + \frac{X_{16}}{16} \tag{22}$$

We repeatedly apply the result on the sum of variances of random variables we quoted from [8] above to get

$$\mathrm{Var}\left(\bar{X}\right) = \mathrm{Var}\left(\frac{X_1}{16}\right) + \mathrm{Var}\left(\frac{X_2}{16}\right) + \ldots + \mathrm{Var}\left(\frac{X_{16}}{16}\right). \tag{23}$$

In [7] Orloff and Bloom state that for a continuous random variable Z, $\mathrm{Var}(aZ+b) = a^2\mathrm{Var}\left(Z\right)$.
Therefore we may rewrite equation 23 as

$$\mathrm{Var}\left(\bar{X}\right) = \frac{1}{16^2}\left(\mathrm{Var}\left(X_1\right) + \mathrm{Var}\left(X_2\right) + \ldots + \mathrm{Var}\left(X_16\right)\right). \tag{24}$$

For this problem, Orloff and Bloom give us that $X_1, X_2, \ldots, X_16$ are identical copies of X, so the variances of $X_1, X_2, \ldots, X_16$ are all equal to the variance of X.
We computed the variance of X above. $\mathrm{Var}\left(X\right) = \frac{3}{80}$.
This implies that

$$\mathrm{Var}\left(\bar{X}\right) = \frac{16}{16^2}\left(\frac{3}{80}\right). \tag{25}$$

Hence the standard deviation of $\bar{X}$ is

$$\sigma_{\bar{X}} = \sqrt{\frac{3}{80\times 16}} \approx 0.0484. \tag{26}$$

2.4 PDF of function of X

Orloff and Bloom ask us to let $Y = X^4$, and ask us to find the pdf of Y.
Let F_y be the cdf of Y. Then:

$$P\left(Y < y\right) = P\left(X^4 < y\right). \tag{27}$$

For $y \neq 0$, equation 27 is true if, and only if

$$P\left(Y < y\right) = P\left(X < \sqrt[4]{y}\right). \tag{28}$$

The cdf of X is x^3. Furthermore, the cdf of X is defined on the interval $[0,1]$ so equation 28 holds if, and only if:

$$P\left(Y < y\right) = x^3\Big|_0^{\sqrt[4]{y}}. \tag{29}$$

We can evaluate equation 29 over the interval $\left[0, \sqrt[4]{y}\right]$; therefore, equation 29 is true if, and only if:

$$P\left(Y < y\right) = \left(\sqrt[4]{y}\right)^3. \tag{30}$$

Equation 29 is true if, and only if, the cdf of Y is $\left(\sqrt[4]{y}\right)^3$.
The pdf is the derivative of the cdf, therefore the pdf of Y is: $\frac{3}{4}y^{\frac{-1}{4}}$

3 Histograms

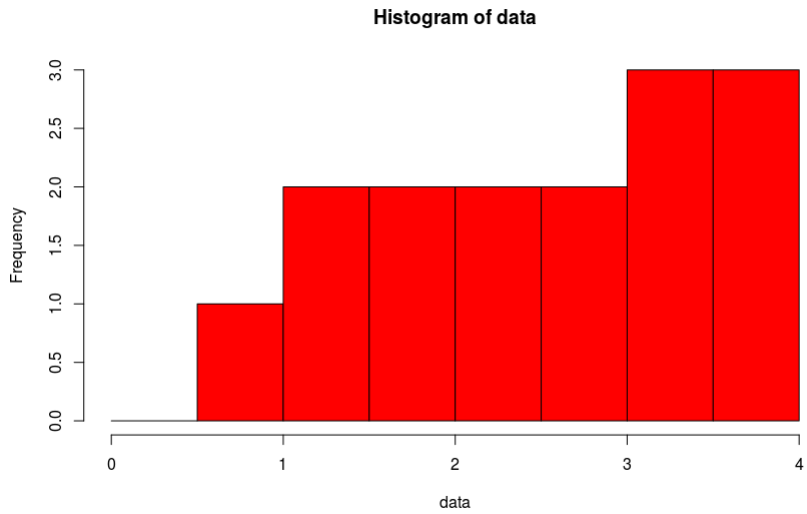
We use R-studio in order to draw the histograms for this list of numbers:
1, 1.2, 1.3, 1.6, 1.6, 2.1, 2.2, 2.6, 2.7, 3.1, 3.2, 3.4, 3.8, 3.9, 3.9
We rely on the examples in [4] in order to create the histograms below.

3.1 Equal bin widths of 0.5

3.1.1 Frequency Histogram

The R code we write to generate the image below is:

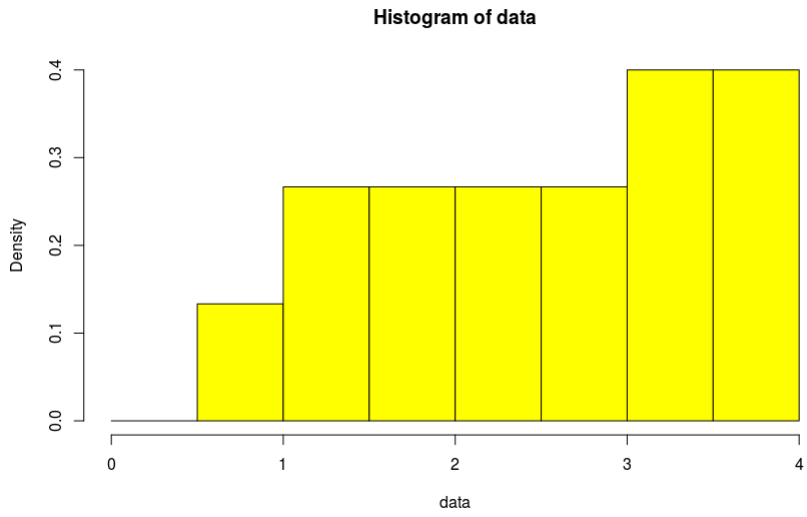
```
binwidth = .5;
data = c(1, 1.2, 1.3, 1.6, 1.6, 2.1, 2.2, 2.6, 2.7, 3.1, 3.2, 3.4, 3.8, 3
bins = seq(0, 4, 0.5)
hist(data, breaks=bins, col='red', freq=TRUE)
```



3.1.2 Density Histogram

The R code to generate the image below is:

```
binwidth = .5;
data = c(1, 1.2, 1.3, 1.6, 1.6, 2.1, 2.2, 2.6, 2.7, 3.1, 3.2, 3.4, 3.8, 3
bins = seq(0, 4, 0.5)
hist(data, breaks=bins, col='yellow', freq=FALSE)
```

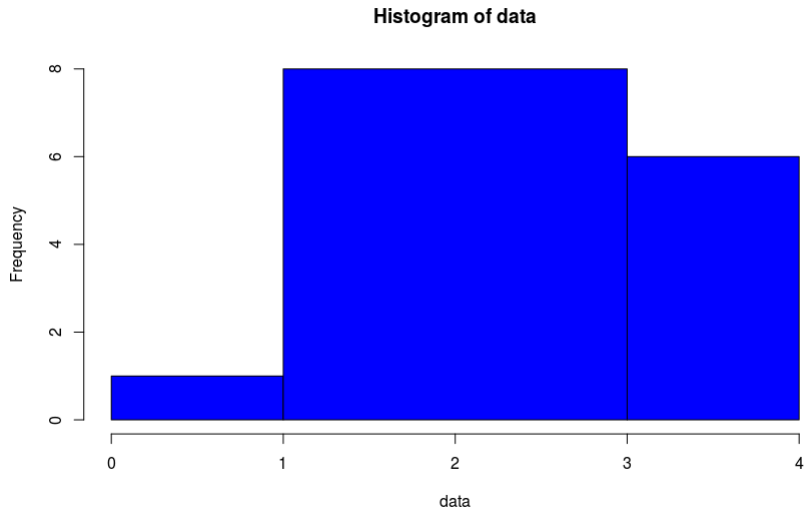


3.2 Unequal bin widths

3.2.1 Frequency Histogram

The R code we write to generate the image below is:

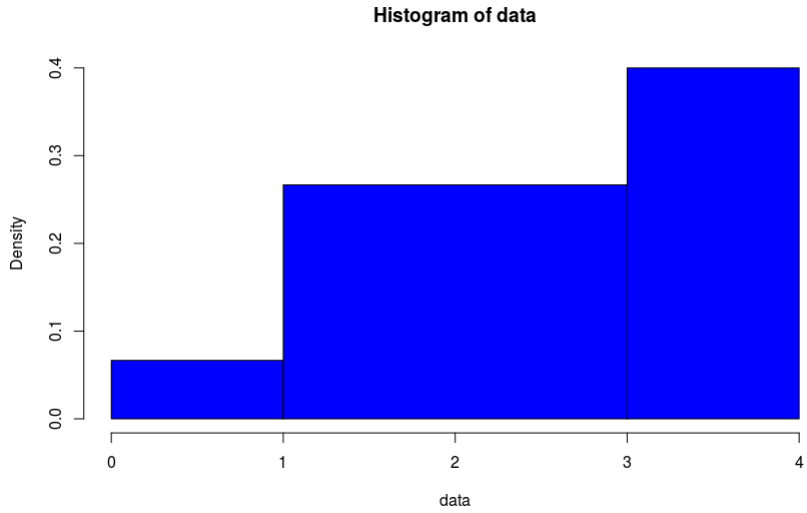
```
binwidth = .5;
data = c(1, 1.2, 1.3, 1.6, 1.6, 2.1, 2.2, 2.6, 2.7, 3.1, 3.2, 3.4, 3.8, 3
bins = c(0,1,3,4)
hist(data, breaks=bins, col='blue', freq=TRUE)
```



3.2.2 Density Histogram

The R code we write to generate the image below is:

```
binwidth = .5;
data = c(1, 1.2, 1.3, 1.6, 1.6, 2.1, 2.2, 2.6, 2.7, 3.1, 3.2, 3.4, 3.8, 3
bins = c(0,1,3,4)
hist(data, breaks=bins, col='blue', freq=FALSE)
```



4 Election

4.1 Central Limit Theorem

The first part of the board question that Orloff and Bloom ask us is a task to carefully write the central limit theorem.
The central limit theorem is:
Suppose $X_1, X_2, \ldots, X_n, \ldots$ are independent identically distributed random variables each having mean μ, and standard deviation σ. For each n let S_n denote the sum and let $\bar{X_n}$ be the average of $X_1, \ldots, X_n$.

$$S_n = X_1 + X_2 + \ldots + X_n = \sum_{i=i}^{n} X_i \tag{31}$$

$$\bar{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{S_n}{n} \tag{32}$$

The properties of mean and variance show: $E(S_n) = n\mu$, $\mathrm{Var}(S_n) = n\sigma^2$, $\sigma_{S_n} = \sqrt{n}\sigma$, $E(\bar{X_n}) = \mu$, $\mathrm{Var}(\bar{X_n}) = \frac{\sigma^2}{n}$, $\sigma_{\bar{X_n}} = \frac{\sigma}{\sqrt{n}}$.

Since they are multiples of each other, S_n and $\bar{X_n}$ have the same standardization

$$Z_n = \frac{S_n - \mu}{\sigma\sqrt{n}} = \frac{\bar{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{33}$$

Central Limit Theorem: for large n,

$$X_n \approx N\left(\mu, \frac{\sigma^2}{n}\right), S_n \approx N\left(n\mu, n\sigma^2\right), Z_n \approx N\left(0, 1\right) \tag{34}$$

4.2 Poll result probability

For the second part of the question, Orloff and Bloom ask us to use the central limit theorem to estimate the probability p that we conduct a poll where we know that 50% of the population supports the candidate, but the poll shows that at least 55% of the population supports a candidate named Erika. Orloff and Bloom write that the population size is 400 people.
We let $\mathcal{E}$ be the discrete random variable that is equal to the number of people that say they will vote for Erika when we conduct the poll.
In order to apply the central limit theorem to estimate p, we must standardize $i\mathcal{E}$ [8], [6].
We model the act of polling one person as a trial of a Bernoulli experiment [8]. $\mathcal{E}$ is then the sum of the number of Bernoulli experiments where the person we poll says she or he will vote for Erika. Then we can apply the central limit theorem because $\mathcal{E}$ is the sum of independent, identically distributed random numbers.
We are given that the size of the population is 400 people, and the fraction of people that vote for Erika is %50. Therefore the expected value, μ of $\mathcal{E}$ is 200.
We use the formula Orloff and Bloom derive in [10] for the variance of a Bernoulli random variable. The probability q of one person we poll saying he or she will vote for Erika is 50%, so according [10], the standard deviation of the Bernoulli random variable that is 1 if the person says she or he will vote for Erika, and 0 otherwise is $(1 - 0.5)(0.5) = 0.25$. We can apply the property of variance Orloff and Bloom show [10] that the variance of the sum of random variables is the sum of the variances of the terms of the sum to get $\text{Var}(\mathcal{E}) = 0.25 \times 400 = 100$. Therefore the standard deviation of $\mathcal{E}$ is 10.
We have the mean and standard deviation of $\mathcal{E}$, so we can standardize $\mathcal{E}$:

$$\frac{\mathcal{E} - 200}{10}. \tag{35}$$

If 55% of the 400 people we poll answer that they will vote for Erika, then this means that $0.55 \times 400 = 220$ people answer they will vote for Erika. So, in this case, $\mathcal{E} = 220$.

We wish to use the central limit theorem to esitmate $P(\mathcal{E} \geq 220)$.
The standardized value of $\mathcal{E}$ is:

$$\frac{220 - 200}{10} = 2. \tag{36}$$

So we estimate $P(\mathcal{E}) \geq 220$ with $P(Z) \geq 2$.
We justify that this estimate is valid: recall that $\mathcal{E}$ is the sum of the outcomes of 400 independent Berounilli trials, and that the central limit theorem allows us to approximate such sums with the standard normal distribution [8].
We know from the rule of thumb [9] that $P(-2 \leq Z \leq 2) \approx 0.95$. Therefore the area under the curve of the probability density function of Z, ϕ is 0.95. We are interested in the value of $P(Z \geq 2)$. So we are interested in the area under the curve where $Z \geq 2$. We know that the total area under the curve of ϕ is 1, and that ϕ is symmetric with respect to the x axis [9]. Therefore $P(Z \geq 2) \approx \frac{1-0.95}{2} \approx 0.025$.
Therefore the probability of our conducting a poll of 400 people where at least 55% of the people answer that they will vote for Erika is 0.025.

4.3 Probability of poll for Ruthi

In this section we entertain the next question that Orloff and Bloom ask for candidate Ruthi. All of the conditions and their consequences that we write about above apply for this question as well.
Orloff and Bloom give us that 25% of the population votes for Ruthi.
We wish to know the probability that, when we conduct a poll of 400 people, less than 20% of the people say that they will vote for Ruthi.
Let the random variable $\mathcal{R}$ be the number of peoplewe poll that say they will vote for Ruthi.
We wish to use the central limit theorem to estimate $P(\mathcal{R} < 20\%)$.
Now the Bernoulli trial where we ask one person who he or she will vote for has a probability of the outcome where the person says she or he will vote for Ruthi of 025%.
of ϕ

References

[1] Alexander Holt David Carlisle Scott Pakin. *The Great, Big List of LaTeXSymbols*. Available at `https://www.rpi.edu/dept/arc/training/latex/LaTeX_symbols.pdf` (2001/2/7).

[2] Michael Dougherty. *Chapter 6 Basic Integration*. Available at `http://faculty.swosu.edu/michael.dougherty/book/chapter06.pdf` (2012/11/20).

[3] StackExchange.com user jamaicanworm. *Nice-looking p-th roots (in the question itself)*. Available at http://faculty.swosu.edu/michael.dougherty/book/chapter06.pdf (2012/03/22).

[4] Jeremy Orloff and Jonathan Bloom. *Code for Studio 3*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/studio-resources/studio3.zip (Spring 2014).

[5] Jeremy Orloff and Jonathan Bloom. *Continuous Expectation and Variance, the Law of Large Numbers, and the Central Limit Theorem 18.05 Spring 2014 Jeremy Orloff and Jonathan Bloom*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/class-slides/MIT18_05S14_class6slides.pdf (Spring 2014).

[6] Jeremy Orloff and Jonathan Bloom. *Continuous Expectation and Variance, the Law of Large Numbers, and the Central Limit Theorem 18.05 Spring 2014 Jeremy Orloff and Jonathan Bloom*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/class-slides/MIT18_05S14_class6_slides.pdf (Spring 2014).

[7] Jeremy Orloff and Jonathan Bloom. *Expectation, Variance and Standard Deviation for Continuous Random Variables Class 6, 18.05, Spring 2014 Jeremy Orloff and Jonathan Bloom*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6a.pdf (Spring 2014).

[8] Jeremy Orloff and Jonathan Bloom. *Expectation, Variance and Standard Deviation for Continuous Random Variables Class 6, 18.05, Spring 2014 Jeremy Orloff and Jonathan Bloom*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6b.pdf (Spring 2014).

[9] Jeremy Orloff and Jonathan Bloom. *Gallery of Continuous Random Variables Class 5, 18.05, Spring 2014 Jeremy Orloff and Jonathan Bloom*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading5c.pdf (Spring 2014).

[10] Jeremy Orloff and Jonathan Bloom. *Variance of Discrete Random Variables Class 5, 18.05, Spring 2014 Jeremy Orloff and Jonathan Bloom*. Available at https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading5a.pdf (Spring 2014).

[11] ShareLatex.com. *Bold, italics and underlining*. Available at `https://www.sharelatex.com/learn/Bold,_italics_and_underlining` (2017).