Biostatistics [SBE304] (Fall 2019)
Tutorial 3

Continuous RVs and their Probability Distributions

R: Descriptive Methods for Assessing Normality & High quality visualizations

Prof. Ayman M. Eldeib   Asem Alaa

Tuesday 8$^{\text{th}}$ October, 2019

1 Tutorial facts

The problems in this tutorial comprises:

(A) Programming Works:

- **>_** **1** programming in-class demos.
- **2** programming homework.
- **1** self-practicing programming works.

(B) Problem Set:

- **5** problems to be solved in-class.
- **7** problems homework.
- **6** self-practicing problems.

Join this GitHub assignment page to create a repository for your submissions: **TBA**

2 Discrete Random Variables and Probability Distributions

2.1 Pre-class reading

1. Lecture notes of "Continuous Random Variables and their Probability Distributions" by Prof. Ayman M. Eldeib

2. Chapter 4 (pp. 66-93) of Montgomery's textbook

3. From Chapter 5 (pp. 95-125) of Montgomery's textbook, read:

   (a) Marginal Probability Distributions.
   (b) Mean and Variance from Joint Distributions.
   (c) Conditional Probability Distributions.
   (d) Independence

2.2 Chapter overview

2.2.1 Continuous RVs

Probabilities assigned to various outcomes in $\mathcal{S}$ in turn determine probabilities associated with the values of any particular rv X. Recall: an rv X is continuous if its set of possible values is uncountable and if

$$P(X = c) = 0 \quad \forall c$$

.

Probability Density Fxn/Probability Distribution, (pdf): $\forall a, b \in \mathbb{R}, a \leq b$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Gives the probability that X takes values between a and b.
The conditions $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) = 1$ are required for any pdf.

Cumulative Distribution Function(cdf):

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

By the continuity arguments for continuous RVs we have that

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a < X < b)$$

Other probabilities can be computed from the cdf $F(x)$:

$$P(X > a) = 1 - F(a)$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

Furthermore, if X is a cont rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which $F'(x)$ exists, $F'(x) = f(x)$.

Median($\tilde{\mu}$): is the 50th percentile st $F(\tilde{\mu}) = .5$. That is half the area under the density curve. For a symmetric curve, this is the point of symmetry.

Expected/Mean Value(μ or $E(X)$): of cont rv with pdf $f(x)$

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

If X is a cont rv with pdf $f(x)$ and $h(X)$ is any function of X then

$$E[h(X)] = \mu = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

Variance: of a cont rv X with pdf $f(x)$ and mean value μ is

$$\sigma_x^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx = E[(X - \mu)^2]$$

Alternatively,

$$V(X) = E(X^2) - [E(X)]^2$$

2.2.2 Continuous Distributions

The Normal Distribution, $X \sim N(\mu, \sigma^2)$ **PDF:** with parameters μ and σ where $-\infty < \mu < \infty$ and $0 < \sigma$

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty$$

We can then easily show that $E(X) = \mu$ and $V(X) = \sigma^2$.

Standard Normal Distribution: The specific case where $\mu = 0$ and $\sigma = 1$. Then

$$\text{pdf}: \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad \text{cdf}: \quad \Phi(z) = \int_{-\infty}^{z} \phi(u)du$$

Standardization: Suppose that $X \sim N(\mu, \sigma^2)$. Then

$$Z = (X - \mu)/\sigma$$

transforms X into standard units. Indeed $Z \sim N(0, 1)$.

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

Independence: If $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent, then $X \pm Y \sim N(\mu_x \pm \mu_y, \sigma_x^2 + \sigma_y^2)$

NOTE: By symmetry of the standard normal distribution, it follows that $\Phi(-z) = 1 - \Phi(z) \quad \forall z \in \mathbb{R}$

Normal Approx to Binomial Dist: Let $X \sim Bin(n, p)$. As long as a binomial histogram is not too skewed, Binomial probabilities can be well approximated by normal curve areas.

$$P(X \leq x) = B(x; n, p) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1 - p)}}\right)$$

As a rule of thumb, the approx is adequate provided that both $np \geq 5$ and $n(1-p) \geq 5$.

| Hypergeometric | $\approx$ | Binomial | $\approx$ | Normal |
| distribution | $\frac{n}{N} < 0.1$ | distribution | $np > 5 \& n(1-p) > 5$ | distribution |

Normal Approx to Poisson Dist: Let $X \sim Poisson(\lambda)$ with $E(X) = \lambda$ and $V(x) = \lambda$, then

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

, is approximately a standard normal random variable. The same continuity correction used for the binomial distribution can also be applied. The approximation is good for $\lambda > 5$

The Exponential Distribution, $X \sim Exp(\lambda)$ Model for lifetime of firms/products/humans **Exponential Distribution:** A cont rv X has exp distribution if its pdf is given by

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0 \quad \lambda > 0$$

$F(x, \lambda)$=	$P(X \leq x) = 1 - e^{\lambda x} \quad x \geq 0$
$E(X)$=	$1/\lambda$
$V(X)$=	$1/\lambda^2$
Memoryless Prop:	$P(X > a + x \mid X > a) = P(X > x)$
	for $x \in D, a > 0$

Note: If Y is an rv distributed as a Poisson $p(y; \lambda)$, then the time between consecutive Poisson events is distributed as an exponential rv with parameter λ

2.2.3 Joint Probability Dist

Joint Probability Mass Fxn: For two discrete rv's X and Y. The joint pmf of (X, Y) is defined $\forall (x, y) \in \mathbb{D}$

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

It must be that $p(x, y) \geq 0$ and $\sum_i \sum_j p(x_i, y_j) = 1$.

Marginal Prob Mass Fxn: of X and of Y, denoted $p_X(x)$ and $p_Y(y)$ respectively,

$$p_X(x) = \sum_{y:p(x,y)>0} p(x, y) \quad \forall x \in \mathbb{D}_1$$

Joint Probability Density Fxn: For two continuous rv's X and Y. The joint pdf of (X, Y) is defined $\forall A \subseteq \mathbb{R}^2$

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

It must be that $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$. Note also that this integration is commutative.

Marginal Prob Density Fxn: of X and of Y, denoted $f_X(x)$ and $f_Y(y)$ respectively,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \forall x \in \mathbb{D}_1$$

Note that if $f(x, y)$ is the joint density of the random vector (X, Y) and $A \in \mathbb{R}^2$ is of the form $A = [a, b] \times [c, d]$ we have that

$$P((X, Y) \in A) = \int_c^d \int_a^b f(x, y) dx dy = \int_a^b \int_c^d f(x, y) dx dy$$

Independence: Two rvs are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad f(x, y) = f_X(x) f_Y(y)$$

Conditional Distribution(discrete): For two discrete rv's X and Y with joint pmf $p(x_i, y_j)$ and marginal X pmf $p_X(x)$, then for any realized value x in the range of X, the conditional mass function of Y, given that $X = x$ is

$$p_{Y|X}(y|x) = \frac{p(x_i, y_j)}{p_X(x)}$$

Conditional Distribution(cont): For two continuous rv's X and Y with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$, then for any realized value x in the range of X, the conditional density function of Y, given that $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

2.2.4 Expected Values, Covariance & Correlation

Expected value: The expected value of a function $h(X, Y)$ of two jointly distributed random variables is

$$E(g(X, Y)) = \sum_{x \in \mathbb{D}_1} \sum_{y \in \mathbb{D}_2} g(x, y) p(x, y)$$

and can be generalized to the continuous case with integrations.//
   Covariance: Measures the strength of the relation btwn 2 RVs, however very

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Shortcut Formula:

$$Cov(X, Y) = E(XY) - \mu_x \mu_y$$

   The defect of the covariance however is that its value depends critically on the units of measurement.
   Correlation: Cov after standardization. Helps interpret Cov.

$$\rho = \rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

Has the property that $Corr(aX + b, cY + d) = Corr(X, Y)$
and that for any rvs X, Y $-1 \leq \rho \leq 1$.

   Note also that ρ is independent of units, the larger $|\rho|$ the stronger the linear association, considered strong linear relationship if $|\rho| \geq 0.8$.

   Caution though: if X and Y are independent then $\rho = 0$ but $\rho = 0$ does not imply that X,Y are independent.
Also that $\rho = 1$ or -1 iff $Y = aX + b$ for some a, b with $a \neq 0$.

2.3 Problem Set

---

1. PROBLEM

   The waiting time for service at a hospital emergency department (in hours) follows a distribution with probability density function $f(x) = 0.5e^{-0.5x}$ for 0 < x. Determine the following:

1. $P(X < 0.5)$

2. $P(X > 2)$

3. Value x (in hours) exceeded with probability 0.05

1. SOLUTION

---

2. PROBLEM

   Suppose that $f(x) = x/8$ for $3 < x < 5$. Determine the following probabilities:

1. $P(X < 4)$

2. $P(X > 3.5)$

3. $P(4 < X < 5)$

4. $P(X < 4.5)$

5. $P(X < 3.5 \, or \, X > 4.5)$

2. SOLUTION

3. Problem

The probability density function of the time you arrive at a terminal (in minutes after 8:00 A.M.) is $f(x) = 0.1e^{-0.1x}$ for $0 < x$. Determine the probability that

1. You arrive by 9:00 A.M.

2. You arrive between 8:15 A.M. and 8:30 A.M.

3. You arrive before 8:40 A.M. on two or more days of five days. Assume that your arrival times on different days are independent.

4. Determine the cumulative distribution function and use the cumulative distribution function to determine the probability that you arrive between 8:15 A.M. and 8:30 A.M.

3. Solution

4. Problem

The time until recharge for a battery in a laptop computer under common conditions is normally distributed with a mean of 260 minutes and a standard deviation of 50 minutes.

1. What is the probability that a battery lasts more than four hours?

2. What are the quartiles (the 25% and 75% values) of battery life?

3. What value of life in minutes is exceeded with 95% probability?

4. Solution

5. Problem

Cholesterol is a fatty substance that is an important part of the outer lining (membrane) of cells in the body of animals. Its normal range for an adult is 120-240 mg/dl. The Food and Nutrition Institute of the Philippines found that the total cholesterol level for Filipino adults has a mean of 159.2 mg/dl and 84.1% of adults have a cholesterol level less than 200 mg/dl (www.fnri.dost.gov.ph/). Suppose that the total cholesterol level is normally distributed.

1. Determine the standard deviation of this distribution.

2. What are the quartiles (the 25% and 75% percentiles) of this distribution?

3. What is the value of the cholesterol level that exceeds 90of the population?

4. An adult is at moderate risk if cholesterol level is more than 1 but less than 2 standard deviations above the mean. What percentage of the population is at moderate risk accord- ing to this criterion?

5. An adult whose cholesterol level is more than 2 standard deviations above the mean is thought to be at high risk. What percentage of the population is at high risk?

6. An adult whose cholesterol level is less than 1 standard deviation below the mean is thought to be at low risk. What percentage of the population is at low risk?

5. Solution

6. Problem

In 2002, the average height of a woman aged 20–74 years was 64 inches, with an increase of approximately 1 inch from 1960 (http://usgovinfo.about.com/od/healthcare). Suppose the height of a woman is normally distributed with a standard deviation of 2 inches.

1. What is the probability that a randomly selected woman in this population is between 58 inches and 70 inches?

2. What are the quartiles of this distribution?

3. Determine the height that is symmetric about the mean that includes 90% of this population.

4. What is the probability that five women selected at random from this population all exceed 68 inches?

6. Solution

7. Problem

Hits to a high-volume Web site are assumed to follow a Poisson distribution with a mean of 10,000 per day. Approximate each of the following:

1. Probability of more than 20,000 hits in a day.

2. Probability of less than 9900 hits in a day.

3. Value such that the probability that the number of hits in a day exceeds the value is 0.01.

4. Expected number of days in a year (365 days) that exceed 10,200 hits.

5. Probability that over a year (365 days), each of the more than 15 days has more than 10,200 hits.

7. Solution

8. Problem

The time between arrivals of taxis at a busy intersection is exponentially distributed with a mean of 10 minutes.

1. What is the probability that you wait longer than 1 hour for a taxi?

2. Suppose that you have already been waiting for 1 hour for a taxi. What is the probability that one arrives within the next 10 minutes?

3. Determine x such that the probability that you wait more than x minutes is 0.10.

4. Determine x such that the probability that you wait less than x minutes is 0.90.

5. Determine x such that the probability that you wait less than x minutes is 0.50.

8. Solution

🏠 9. PROBLEM

The distance between major cracks in a highway follows an exponential distribution with a mean of 5 miles.

1. What is the probability that there are no major cracks in a 10-mile stretch of the highway?

2. What is the probability that there are two major cracks in a 10-mile stretch of the highway?

3. What is the standard deviation of the distance between major cracks?

4. What is the probability that the first major crack occurs between 12 and 15 miles of the start of inspection?

5. What is the probability that there are no major cracks in two separate 5-mile stretches of the highway?

6. Given that there are no cracks in the first 5 miles inspected, what is the probability that there are no major cracks in the next 10 miles inspected?

9. SOLUTION

💾 10. PROBLEM

According to an analysis of chocolate bars in chocolate factory, the mean number of insect fragments was 14.4 in 225 grams. Assume that the number of fragments follows a Poisson distribution.

1. What is the mean number of grams of chocolate until a fragment is detected?

2. What is the probability that there are no fragments in a 28.35-gram (one-ounce) chocolate bar?

3. Suppose you consume seven one-ounce (28.35-gram) bars this week. What is the probability of no insect fragments?

10. SOLUTION

11. PROBLEM

The length of stay at a specific emergency department in a hospital in Phoenix, Arizona, had a mean of 4.6 hours. Assume that the length of stay is exponentially distributed.

1. What is the standard deviation of the length of stay?

2. What is the probability of a length of stay of more than 10 hours?

3. What length of stay is exceeded by 25% of the visits?

11. SOLUTION

12. PROBLEM

Calls to a telephone system follow a Poisson process with a mean of five calls per minute.

1. What is the name applied to the distribution and parameter values of the time until the 10th call?

2. What is the mean time until the 10th call?

3. What is the mean time between the 9th and 10th calls?

4. What is the probability that exactly four calls occur within 1 minute?

5. If 10 separate 1-minute intervals are chosen, what is the probability that all intervals contain more than two calls?

12. SOLUTION

13. PROBLEM

The time between arrivals of customers at an automatic teller machine is an exponential random variable with a mean of 5 minutes.

1. What is the probability that more than three customers arrive in 10 minutes?

2. What is the probability that the time until the fifth customer arrives is less than 15 minutes?

13. SOLUTION

14. PROBLEM

Determine the value of c such that the function $f(x, y) = cxy$ for $0 < x < 3$ and $0 < y < 3$ satisfies the properties of a joint probability density function. Determine the following:

1. $P(X < 2, Y < 3)$

2. $P(X < 2.5)$

3. $P(1 < Y < 2.5)$

4. $P(X > 1.8, 1 < Y < 2.5)$

5. $E(X)$

6. $P(X < 0, Y < 4)$

7. Marginal probability distribution of X

14. SOLUTION

15. PROBLEM

Determine the value of c that makes the function $f(x, y) = ce^{-2x-3y}$ a joint probability density function over the range $0 < x$ and $x < y$. Determine the following:

1. $P(X < 1, Y < 2)$

2. $P(1 < X < 2)$

3. $P(Y > 3)$

4. $P(X < 2, Y < 2)$

5. $E(X)$

6. $E(Y)$

7. Marginal probability distribution of X

15. SOLUTION

16. PROBLEM

An article in Health Economics ["Estimation of the Transition Matrix of a Discrete-Time Markov Chain" (2002, Vol. 11, pp. 33–42)] considered the changes in CD4 white blood cell counts from one month to the next. The CD4 count is an important clinical measure to determine the severity of HIV infections. The CD4 count was grouped into three distinct categories: $0-49, 50-74$, and ≥ 75. Let X and Y denote the (category minimum) CD4 count at a month and the following month, respectively. The conditional probabilities for Y given values for X were provided by a transition probability matrix shown in the following table.

X	Y		
	0	50	75
0	0.9819	0.0122	0.0059
50	0.1766	0.7517	0.0717
75	0.0237	0.0933	0.8830

This table is interpreted as follows. For example, $P(Y = 50|X = 75) = 0.0933$. Suppose also that the probability distribution for X is $P(X = 75) = 0.9$, $P(X = 50) = 0.08$, $P(X = 0) = 0.02$. Determine the following:

1. $P(Y \leq 50|X = 50)$

2. $P(X = 0, Y = 75)$

3. $E(Y|X = 50)$

4. $f_{XY}(x, y)$

5. $f_Y(y)$

6. Are X and Y independent?

16. SOLUTION

The systolic and diastolic blood pressure values (mmHg) are the pressures when the heart muscle contracts and relaxes (denoted as Y and X, respectively). Over a collection of individuals, the distribution of diastolic pressure is normal with mean 73 and standard deviation 8. The systolic pressure is conditionally normally distributed with mean $1.6x$ when $X = x$ and standard deviation of 10. Determine the following:

1. Conditional probability density function of Y given $X = 73$.

2. $P(Y < 115|X = 73)$

3. $E(Y|X = 73)$

4. Recognize the distribution $f_{XY}(x, y)$ and identify the mean and variance of Y.

17. SOLUTION

---

18. PROBLEM

Suppose that X and Y have a bivariate normal distribution with $\sigma_X = 0.04$, $\sigma_Y = 0.08$, $\mu_X = 3.00$, $\mu_Y = 7.70$, and $\rho = 0$.

Determine the following:

1. $P(2.95 < X < 3.05)$

2. $P(7.60 < Y < 7.80)$

3. $P(2.95 < X < 3.05, 7.60 < Y < 7.80)$

18. SOLUTION

---

3   ℝ Programming Language

3.1   Quantile Quantile Plots

"In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions." *(Wikipedia)*

3.2   ℝ Programming Problems

---

1. PROBLEM

For these exercises, we will be using the following dataset:
```
library(downloader)
url <- "https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/ext"
"data/femaleControlsPopulation.csv"
filename <- basename(url)
download(url, destfile=filename)
x <- unlist( read.csv(filename) )
```
Here `x` represents the weights for the entire population.

1. What is the average of these weights?

2. After setting the seed at 1, `set.seed(1)` take a random sample of size 5. What is the absolute value (use abs) of the difference between the average of the sample and the average of all the values?

3. After setting the seed at 5, `set.seed(5)` take a random sample of size 5. What is the absolute value of the difference between the average of the sample and the average of all the values?

4. Why are the answers from 2 and 3 different?

   (a) Because we made a coding mistake.
   (b) Because the average of the `x` is random.
   (c) Because the average of the samples is a random variable.
   (d) All of the above.

5. Set the seed at 1, then using a for-loop take a random sample of 5 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of `x` ?

6. We are now going to increase the number of times we redo the sample from 1,000 to 10,000. Set the seed at 1, then using a for-loop take a random sample of 5 mice 10,000 times. Save these averages. What percent of these 10,000 averages are more than 1 ounce away from the average of `x` ?

7. Note that the answers to 4 and 5 barely changed. This is expected. The way we think about the random value distributions is as the distribution of the list of values obtained if we repeated the experiment an infinite number of times. On a computer, we can't perform an infinite number of iterations so instead, for our examples, we consider 1,000 to be large enough, thus 10,000 is as well. Now if instead we change the sample size, then we change the random variable and thus its distribution. Set the seed at 1, then using a for-loop take a random sample of 50 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of `x` ?

>_   **2. PROBLEM**

Quantile-Quantile Plots in Action   Using the standard `trees` dataset. Hints: check the documentation of `qqnorm` , `qqline` , `qqplot`

1. Check the normality of trees height (`trees$Height`).

2. Compare the distribution of trees Girth (`trees$Girth`) and its Volume (`trees$Volume`).

⟨/⟩   **3. PROBLEM**

Learning R's ggplot2   Choose one of the excellent introductory courses/tutorials from DataCamp to learn generating beautiful and high quality data visualizations:

1. Introduction to Data Visualization with ggplot2

2. Data Visualization with ggplot2

   (a) Part 1
   (b) Part 2
   (c) Part 3

⟨/⟩   **4. PROBLEM**

Let X and Y represent the concentration and viscosity of a chemical product. Suppose that X and Y have a bivariate normal distribution with $\sigma_X = 4$, $\sigma_Y = 1$, $\mu_X = 2$, and $\mu_Y = 1$. Draw a rough contour plot of the joint probability density function for each of the following values of ρ:

1. $\rho = 0$

2. $\rho = 0.8$

3. $\rho = -0.8$