Distributions: Normal and Poisson

Chapter 3, Lab 2: Solutions

OpenIntro Biostatistics

Topics

- Normal distribution
- Poisson distribution

This lab discusses two additional distributions: the normal and Poisson. Working with normal probabilities and evaluating the normal approximation will be especially important in later units on inference. The Poisson distribution is useful for estimating the rate that events occur in a large population over a unit of time.

The material in this lab corresponds to Sections 3.3 and 3.4 of *OpenIntro Biostatistics*.

Normal distribution

A normal distribution curve is characterized by two parameters: the mean (μ) and standard deviation (σ). The normal distribution with mean 0 and standard deviation 1 is referred to as the standard normal distribution; a random variable following the standard normal distribution is typically denoted as Z.

The z-score of an observation quantifies how far the observation is from the mean, in units of standard deviation(s). The z-score for an observation x from a random variable $X \sim \text{N}(\mu, \sigma)$ equals

$$z = \frac{x - \mu}{\sigma}.$$

In the reverse direction, the value x from $X \sim \text{N}(\mu, \sigma)$ associated with a particular z-score equals

$$x = \mu + \sigma z.$$

Normal probabilities are calculated in R with the use of `pnorm`.

Unlike the binomial, the normal distribution is continuous. Since the probability of a continuous random variable equaling some exact value is always 0, it is valid to state that for a continuous random variable X, $P(X \leq x) = P(X < x)$ and $P(X \geq x) = P(X > x)$.

The following code shows how to calculate $P(X \leq 105)$ and $P(X > 105)$ for $X \sim \text{N}(100, 5)$. When values of μ and σ are not specified, R assumes that $\mu = 0$ and $\sigma = 1$.

```
#probability X is less than (or equal to) 105
pnorm(105, 100, 5)
```

```
## [1] 0.8413447
```

```r
#probability X is greater than 105
pnorm(105, 100, 5, lower.tail = FALSE)
```

```
## [1] 0.1586553
```

```r
#probability Z is less than (or equal to) 1
pnorm(1)
```

```
## [1] 0.8413447
```

```r
#probability Z is greater than 1
pnorm(1, lower.tail = FALSE)
```

```
## [1] 0.1586553
```

To identify the observation corresponding to a particular probability, use `qnorm`. The following code shows how to identify the X or Z value where there is 0.841 area to the left (and 0.159 area to the right). The values differ slightly from the ones above due to rounding.

```r
#identify X value
qnorm(0.841, 100, 5)
```

```
## [1] 104.9929
```

```r
qnorm(0.159, 100, 5, lower.tail = FALSE)
```

```
## [1] 104.9929
```

```r
#identify Z value
qnorm(0.841)
```

```
## [1] 0.9985763
```

```r
qnorm(0.159, lower.tail = FALSE)
```

```
## [1] 0.9985763
```

1. In the last decade, the average age of a mother at childbirth is 26.4 years, with standard deviation 5.8 years. The distribution of age at childbirth is approximately normal.

   a) What age at childbirth puts a woman in the upper 2.5% of the age distribution? In other words, what is the 97.5 percentile of this age distribution?

   An age of about 38 (37.8) years puts a woman in the upper 2.5% of the age distribution

```
qnorm(0.975, mean = 26.4, sd = 5.8)
```

```
## [1] 37.76779
```

   b) What proportion of women who give birth are 21 years of age or older?

   $P(X \geq 21) = 0.824$; 82.4% of women are 21 years of age or older.

```
pnorm(21, mean = 26.4, sd = 5.8, lower.tail = FALSE)
```

```
## [1] 0.8240821
```

2. Suppose a mild hypertensive is defined as a person whose diastolic blood pressure (DBP) is between 90 and 100 mm Hg, inclusive. Assume that in the population of 35-44 year old men, mean DBP is 80 mm Hg, with variance 144. What is the probability that a randomly selected male from this population is a mild hypertensive, assuming that DBP is normally distributed?

   Find $P(90 \leq X \leq 100) = P(X \leq 100) - P(X \leq 90)$. The probability that a randomly selected male from this population is a mild hypertensive is 0.154.

```
pnorm(100, mean = 80, sd = 12) - pnorm(90, mean = 80, sd = 12)
```

```
## [1] 0.154538
```

3. People are classified as hypertensive if their systolic blood pressure (SBP) is higher than a specified level for their age group. For ages 1-14, SBP has a mean of 105.0 and standard deviation 5.0, with hypertension level 115.0. For ages 15-44, SBP has a mean of 125.0 and standard deviation 10.0, with hypertension level 140.0. Assume SBP is normally distributed.

   Define a family as a group of two people in age group 1-14 and two people in age group 15-44. A family is classified as hypertensive if at least one adult and at least one child are hypertensive.[1]

   a) What proportion of 1- to 14-year-olds are hypertensive?

   $P(X \geq 115.0) = 0.0228$; 2.28% of 1- to 14-year-olds are hypertensive.

```
pnorm(115, mean = 105, sd = 5, lower.tail = FALSE)
```

```
## [1] 0.02275013
```

   b) What proportion of 15- to 44-year-olds are hypertensive?

   $P(Y \geq 140.0) = 0.0668$; 6.68% of 15- to 44-year-olds are hypertensive.

```
pnorm(140, mean = 125, sd = 10, lower.tail = FALSE)
```

[1] Problem from Rosner, *Fundamentals of Biostatistics*, 7^{th} edition, pp. 138-139

```
## [1] 0.0668072
```

c) What is the probability that a family is hypertensive? Assume that the hypertensive status of different members of a family are independent random variables. (Admittedly, this assumption is highly unrealistic.)

Let C be a binomial random variable modeling the number of children in a family that are hypertensive, and A be a binomial random variable modeling the number of hypertensive adults in a family; $C \sim \text{Bin}(2, 0.0228)$, $A \sim \text{Bin}(2, 0.0668)$.

A family is considered hypertensive if at least one adult and at least one child are hypertensive. Let H represent the event that a family is hypertensive. Assume that C and A are independent:

$$P(H) = P(C \geq 1) \times P(A \geq 1) = 0.0058$$

The probablity that a family is hypertensive is 0.0058.

```
pbinom(0, 2, 0.0228, lower.tail = FALSE) * pbinom(0, 2, 0.0668, lower.tail = FALSE)
```

```
## [1] 0.005821551
```

d) Consider a community of 1,000 families. What is the probability that between one and five families (inclusive) are hypertensive?

Let K be a binomial random variable modeling the number of hypertensive families in the community. $P(1 \leq K \leq 5) = P(K \leq 5) - P(K = 0) = 0.475$. The probability that between one and five families are hypertensive is 0.475.

```
pbinom(5, 1000, 0.0058) - dbinom(0, 1000, 0.0058)
```

```
## [1] 0.4749525
```

Poisson distribution

The Poisson distribution is characterized by a single parameter, λ, which expresses the rate of event occurrences per unit time; $X \sim \text{Pois}(\lambda)$. The probability that exactly k events occur in t units of time is

$$P(X = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$

For a Poisson random variable, $E(X) = \lambda$ and $\text{Var}(X) = \lambda$.

The following code shows how to calculate $P(X = 5)$, $P(X \leq 5)$, and $P(X > 5)$ for $X \sim \text{Pois}(3)$.

```
#probability X equals 5
dpois(5, 3)
```

```
## [1] 0.1008188
```

```
#probability X is less than or equal to 5
ppois(5, 3)
```

```
## [1] 0.9160821
```

```
#probability X is greater than 5
ppois(5, 3, lower.tail = FALSE)
```

```
## [1] 0.08391794
```

4. The rate of trisomy 21 (Down syndrome) birth defects is about 1/800 live births per year. In 2012, there were 8,011 live births in the city of Boston. Assume that the number of live births has remained constant over the past few years (2012-2017).

   a) What is the expected number of children born with Down syndrome in Boston in 2017?

   The expected number of children born with Down syndrome in Boston in 2017, assuming that the number of live births has remained constant, is (1/800)(8011) = 10.01.

   ```
   lambda = 1/800; n = 8011
   n*lambda
   ```

   ```
   ## [1] 10.01375
   ```

   b) What is the probability of 12 or more Down syndrome births in 2017?

   Since the expected value of a Poisson distribution is the same as the rate λ, the expected number of births with trisomy 21 is 10.01. The probability of 12 or more Down syndrome births in 2017 is $P(X \geq 12)$, where $X \sim \text{Pois}(10.01)$. The probability that 12 or more Down syndrome births occur is 0.304.

   ```
   ppois(11, lambda = 10.01, lower.tail = FALSE)
   ```

   ```
   ## [1] 0.3043618
   ```

   c) In 2012, 23% of the 8,011 live births were among women age 35 or older. Is this enough information to recalculate parts a) and b) for women age 35 or older? Explain your answer.

No, this is not enough information, because it is not sufficient to only adjust for the smaller number of births from women age 35 or older. Women in this age group are more likely to have a Down syndrome birth; it is not reasonable to assume that the rate of trisomy 21 birth defects for this age group is 1/800 live births per year.

d) Rates in populations are sometimes expressed as a rate per 1,000 or 10,000. Could the following statement be true, or is it clearly a contradiction? Explain your answer.

In 2012, there were 44.5 births per 1,000 female Boston residents, ages 15-44. There was no significant change in the Boston birth rate between 2008 and 2012. In 2012, the rates of births to Black and Latino women in the same age range were 64.7 and 66.3, respectively.

The statement could be true, even if the rates of births to Black and Latino women are higher than the overall average rate; this suggests that the rate of births to other groups of women, such as Caucasian or Asian women, are lower than the overall average rate of 44.5 births per 1,000 female residents ages 15-44.

5. Hemophilia is a sex-linked bleeding disorder that slows the blood clotting process. In severe cases of hemophilia, continued bleeding occurs after minor trauma or even in the absence of injury. The annual rate of males born with hemophilia in the US is approximately 1 per 5,000 male births. In the United States, there are approximately 4,000,000 births per year. Assume that there are equal numbers of males and females born each year.

a) What is the probability that at most 380 newborns in a year are born with hemophilia?

Assuming that there are equal numbers of males and females born each year, there are approximately 2,000,000 male births per year in the United States; thus, the rate λ for a population of 2,000,000 is $(1/5,000)(2,000,000) = 400$. Since the expected value of a Poisson distribution is the same as the rate λ, the expected number of male births with hemophilia is 400. The probability that at most 380 newborn males have hemophilia is $P(X \leq 380)$, where $X \sim \text{Pois}(400)$. The probability that at most 380 newborns in a year are born with hemophilia is 0.165.

```
rate = (1/5000)*(4000000/2)
ppois(380, lambda = rate)
```

```
## [1] 0.164859
```

b) What is the probability that 450 or more newborns in a year are born with hemophilia?

The probability that 450 or more newborns in a year are born with hemophilia is 0.007.

```
ppois(449, lambda = rate, lower.tail = FALSE)
```

```
## [1] 0.007454327
```

c) Consider a hypothetical country in which there are approximately 1.5 million births per year. If the incidence rate of hemophilia is equal to that in the US, as well as the sex ratio at birth, how many newborns are expected to have hemophilia over five years, and with what standard deviation?

The number of male births per year in the hypothetical country is $(1/2)(1,500,000) = 750,000$. If in this hypothetical country, the annual rate of hemophilia in male births is 1 per 5,000, then the expected number of male births with hemophilia in one year

is $(1/5,000)(750,000) = 150$; the rate λ for one year is 150. Over five years, the rate of hemophilia births is $(150)(5) = 750$. Since the rate λ for a Poisson distribution is both the mean and the variance, the expected number of hemophilia births over five years is 750, with standard deviation $\sqrt{750} = 27.39$.

```
new.yearly.rate = (1/5000)*(1500000/2); new.yearly.rate
```

```
## [1] 150
```

```
years = 5
five.year.rate = new.yearly.rate*years; five.year.rate
```

```
## [1] 750
```

```
sqrt(five.year.rate)
```

```
## [1] 27.38613
```

6. The US Centers for Disease Control (CDC) has been monitoring the rate of deaths from opioid overdoses for at least the last 15 years. As the CDC notes on its website, "Of the 22,767 deaths relating to prescription drug overdose in 2013, 16,325 (71.3%) involved opioid painkillers."[2] This statistic is simply a count of the number of deaths, and does not show a rate per unit of the population, which is more useful for monitoring this phenomenon in different communities. The latest statistic shows that among non-Hispanic whites, the rate of opioid-related deaths has risen to 6.8 deaths per year per 100,000 non-Hispanic whites.

   In 2014-2015, the population of Essex County, MA, was approximately 769,000, of whom 73% are non-Hispanic white. The county health department has asked you to investigate the incidence of opioid-related deaths.

   a) In 2014, Essex County reported 146 overdose fatalities from opioids. Assume that all these deaths occurred in the non-Hispanic white members of the population. What is the probability of 146 or more such events in a given year, assuming that the incidence rate of opioid deaths in Essex County is the same as the national rate?

   The probability of 146 or more overdose fatalities in a given year is 2.62×10^{-40}.

```
non.hispanic.whites = 769000 * 0.73
lambda.essex = (6.8/100000) * non.hispanic.whites
ppois(145, lambda.essex, lower.tail = FALSE)
```

```
## [1] 2.621073e-40
```

   b) What was the observed rate of opioid-related deaths in Essex County in 2014, stated in terms of deaths per 100,000 non-Hispanic white members of the population? Continue to assume that all of the observed deaths were among non-Hispanic whites.

   The observed rate is approximately 26 deaths per year per 100,000 non-Hispanic whites, which is much higher than the national rate reported by the CDC.

```
non.hispanic.whites = 769000 * 0.73
obs.rate = (146/non.hispanic.whites) * 100000
obs.rate
```

[2] http://www.cdc.gov/drugoverdose/data/overdose.html

```
## [1] 26.0078
```

c) In 2015, Essex County reported 165 opioid-related deaths in its non-Hispanic white population. Using the rate from part b), calculate the probability of 165 or more such events.

Using the observed rate from part b), find $P(X \geq 165) = P(X > 164)$. Since the population of Essex County did not change from 2014 to 2015, λ for this calculation is simply 146, which is already a rate in terms of deaths per year for non-Hispanic whites, for the population size of Essex County.

The probability of 165 or more opioid-related deaths is around 0.065, or 6.5%.

```
#show that the observed rate lambda is 146
obs.lambda.essex = (146/non.hispanic.whites) * non.hispanic.whites
obs.lambda.essex
```

```
## [1] 146
```

```
ppois(164, obs.lambda.essex, lower.tail = FALSE)
```

```
## [1] 0.0650307
```

d) Assess whether the findings from parts a) through c) are indicative of an opioid overdose crisis in Essex County.

The probability of 165 or more opioid-related deaths occuring in Essex County in 2015 is still small, even after adjusting the rate by using the number of observed overdose fatalities in 2014. The rate of opioid-related deaths in 2014 for Essex County was already far greater than the national rate, and even more overdose fatalities occurred in 2015. The findings are indicative of a rapid increase in opioid misuse in Essex County, on a scale that is extreme relative to the national rate.