Statistical_Inference

Laura Bennett

June 8, 2017

Statistical Inference for Data Science: Chapter 6 Exercises

1. Your friend claims that changing the font to comic sans will result in more ad revenue on your web sites. When presented in random order, 9 pages out of 10 had more revenue when the font was set to comic sans. If it was really a coin flip for these 10 sites, what's the probability of getting 9 or 10 our of 10 with more revenue for the new font?

- **In this case, we may model the situation with a Binomial distribution, as there are only two possible outcomes for the sites: they will either have increased revenue or they will not. If we are assuming that there is equal probability of either case, than the probability of increased revenue is simply 0.5. Additionally, we are looking for the probability of getting 9 OR 10 sites with increased revenue; thus, we must find the probability of 9 and the probability of 10 sites getting increased revenue and add them.**

```
r   sum(dbinom(9:10,size=10, prob=0.5))
```

```
## [1] 0.01074219
```

2. A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, whats the probability of fewer than 5 documentation errors?

- **In the above case, we are examining a normal distribution s.t.** $X \sim N(11, 2)$**. We may calculate the probability of** $X < 5$ **to be:**

```
r   pnorm(4, mean=11, sd=2)
```

```
## [1] 0.0002326291
```

- **Note that 4 or fewer errors is nearly three standard deviations away from the mean; it makes sense to have a low probability.**

3. The number of search entries entered at a web site is Poisson at a rate of 9 searches per minute. The site is monitored for 5 minutes. What is the probability of 40 or fewer searches in that time frame?

- **For this problem, we have** $\lambda = 9$ **and** $t = 5$**; we may calculate the probability of 40 or fewer searches in this period as:**

```
r   ppois(40,lambda=9*5)
```

```
## [1] 0.2555451
```

4. Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What's the probability that a given day has fewer than 93 hits per day expressed as a percentage to the nearest percentage point?

- **This is a normal distribution of** $X \sim N(\mu = 100, \sigma^2 = 100)$**.**

```
r   pnorm(93,mean=100,sd=10)
```

```
## [1] 0.2419637
```

5. Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What number of web hits per day represents the number so that only 5% of the days have more hits?

- **In this case, it is helpful to remember that '100 hits per day' is the mean s.t. 50% of the days will have less than 100 hits per day and 50% will have more hits per day. In this case, we want to see what number of hits per day represents the 95th quantile, i.e. only 5% of the days will have scored greater than that number of web hits per day.**

```
r   qnorm(0.95, mean=100, sd=10)
```

```
## [1] 116.4485
```

6. Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. Imagine taking a random sample of 50 days. What number of web hits would be the point so that only 5% of averages of 50 days of web traffic have more hits?

- **Since we are taking a sample of the normally distributed data set, we may assume that the sample itself will be normally distributed and centered around 100. The new standard deviation, however, must now be calculated as $S = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}}$. We may then use the same technique as the previous problem to find the solution.**

```
r   qnorm(0.95,mean=100,sd=(10/sqrt(50)))
```

```
## [1] 102.3262
```

7. You don't believe that your friend can discern good wine from cheap. Assuming that you're right, a blind test where you randomize 6 paired varieties (Merlot, Chianti, ...) of cheap and expensive wines. What is the chance that she gets 5 or 6 right?

- **If we are assuming that she is guessing which is the expensive wine, we may assume that this is a binomial distribution with the probability of a successful guess being 50%. Therefore, the probability that she gets either five or six correct will be the sum of each individual probability:**

```
r   sum(dbinom(5:6,size=6, prob=0.5))
```

```
## [1] 0.109375
```

8. The number of web hits to a site is Poisson with mean 16.5 per day. What is the probability of getting 20 or fewer in 2 days?

- **This is a poisson distribution of $X \sim P(\lambda = 16.5, t = 2)$. The probability of getting 20 or fewer hits in a period of 2 days is then:**

```
r   ppois(20,lambda=16.5*2)
```

```
## [1] 0.01045391
```