Chapter 4 Textbook exercises

Solutions to even-numbered questions
Statistics and statistical programming
Northwestern University
MTS 525

Aaron Shaw

October 7, 2020

Contents

All exercises taken from the *OpenIntro Statistics* textbook, 4^{th} edition, Chapter 4.

4.4 Triathlons

(a) Let M denote the finishing times of *Men, Ages 30 - 34* and W denote the finishing times of *Women, Ages 25 - 29.* Then,

$$M \sim N(\mu = 4313, \sigma = 583)$$
$$W \sim N(\mu = 5261, \sigma = 807)$$

(b) Recall that Z-scores are a standardization: for a given value of a random variable you subtract the mean of the corresponding distribution from the value and divide by the standard deviation. The formula notation is given in the *OpenIntro* textbook.

Let's let R calculate it for us:

```r
## Mary
(5513 - 5261) / 807
```

```
## [1] 0.3122677
```

```r
## Leo:
(4948 - 4313) / 583
```

```
## [1] 1.089194
```

Since the Z score tells you how many standard deviation units above/below the mean each value is, we can see that Mary finished 0.31 standard deviations above the mean in her category while Leo finished 1.09 standard deviations above the mean in his.

(c) Mary finished in a much faster time with respect to her reference group. Her time was fewer standard deviation units above the mean, implying that a larger proportion of the distribution had higher (slower) race times.

(d) Note that the question is asking about the area under the distribution to the right (greater than) of Leo's race time. Using the Z-score table (Appendix C.1) in the book, we can see that Leo finished *faster* than approximately $1 - 0.86 = .14$ or 14% of his reference group. This corresponds the probability $P(Z > 1.09)$ for a normal distribution. You could also use R to calculate this (note that the *OpenIntro* reading introduced the `pnorm()` function on p.136):

```
1-pnorm(1.09)
```

```
## [1] 0.1378566
```

(e) Again, this is about calculating the area under the distribution ot the right (greater than) Mary's race time. Mary finished *faster* than approximately $1-0.62 = .38$ or 38% of her category. This corresponds to the probability $P(Z > 0.31)$ for a normal distribution. Again, here's how you could find that using R:

```
1-pnorm(0.31)
```

```
## [1] 0.3782805
```

(f) The answer for part b would not change as standardized values (Z-scores) can be computed for any distribution. However, the interpretation and percentile calculations (parts c-e) *would* be different because they all presume a normal distribution.

4.6 More triathlons

(a) The fastest 5% are the 5^{th} percentile of the distribution. Using the Appendix C.1 table again, the Z score corresponding to the 5^{th} percentile of the normal distribution is approximately -1.65. You can find this value more precisely in R using the `qnorm()` function (more on this in the Week 5 R tutorial):

```
qnorm(.05)
```

```
## [1] -1.644854
```

Once you have that, you can plug it into the Z score formula and calculate the cutoff time (x):

$$Z = -1.64 = \frac{x - 4313}{583} \rightarrow x = -1.64 \times 583 + 4313 = 3357 \ seconds$$

Note that the solution there is in seconds. If you divide that by 60 it looks like the fastest 5% of males in this age group finished in a little bit less than 56 minutes *or less*.

(b) The slowest 10% are in the 90^{th} percentile of the distribution. The Z score corresponding to the 90^{th} percentile of the normal distribution is approximately 1.28. Again, here's that calculation in R:

```
qnorm(.9)
```

```
## [1] 1.281552
```

Then put it all together again to calculate the cutoff:

$$Z = 1.28 = \frac{x - 5261}{807} \rightarrow x = 1.28 \times 807 + 5261 = 6294 \ seconds$$

Divide that by 60 and it looks like the slowest 10% of females in this age group finished in about 1 hour 45 minutes *or more*.

4.22 Arachnophobia

This question focuses on applying the knowledge from section 4.3 of the textbook on binomial distributions. Our old friend the binomial coefficient comes in quite handy…

(a) Recall from the birthday problems that a binomial probability of “at least one” successful trial can also be thought of as “one minus the probability of none.” With this in hand, you can start to plug values into the formula for the probability of observing k successess out of n independent binomial trials given on p. 150.

$$P(\mathit{at\ least\ 1\ arachnophobe}) = 1 - P(\mathit{none})$$

$$1 - P(\mathit{none}) = 1 - \binom{10}{0}0.07^0(1-0.07)^{10-0}$$

Let’s let R handle the arithmetic:

```
1-(choose(10,0)*1*(.93^10))
```

```
## [1] 0.5160177
```

(b) This one just requires you to plug a different value for n into the same formula:

$$P(2\ \mathit{arachnophobes}) = \binom{10}{2}0.07^2(1-0.07)^{(10-2)}$$

```
choose(10,2)*0.07^2*0.93^8
```

```
## [1] 0.1233878
```

(c) You can think of the probability of “at most one” success in a binomial trial as equal to the sum of the probability of two potential outcomes: zero or one.

$$P(\leq 1\ \mathit{arachnophobes}) = P(\mathit{none}) + P(\mathit{one})$$

Off to the races with our same formula again:

$$\binom{10}{0}0.07^0 0.93^{10} + \binom{10}{1}0.07^1 0.93^9$$

And R can solve that quickly:

```
(choose(10,0)*1*(.93^10))+(choose(10,1)*0.07*(0.93^9))
```

```
## [1] 0.8482701
```

(d) The question asks us to calculate whether random assignment to tents is likely to ensure ≤ 1 *arachnophobe* per tent. We can think about this as a slight twist on the result we calculated for part c above. Specifically, the answer to part c is the complementary probability of the outcome we’re looking to avoid in this case (more than 1 arachnophobe per tent). In more formal notation:

$$P(> 1\ \mathit{arachnophobe}) = 1 - P(\leq 1\ \mathit{arachnophobe})$$
$$P(> 1\ \mathit{arachnophobe}) = 1 - 0.84 = 0.16 = 16\%$$

That covers the *probability* of multiple arachnophobes per tent, but as to whether or not it seems “reasonable” to randomly assign the teenagers to tents given this probability, the 16% result cannot answer that part of the question. Making a decision based on a probability is an entirely separate issue! On the one hand, the probability of a bad outcome is not *huge*, but the decision should really depend on how heavily the counselor weighs the negative potential outcome given a 16% chance of having multiple arachnophobic campers in one of the tents. The question makes it sound like the counselor “wants to make sure” there’s not a critical mass of arachnophobes in any one tent, so a

16% probability of failure implies that they should *not* use random assignment. Indeed, if the camp counselor has taken a statistics course, they might consider *any* probability of failure greater than 5% as unacceptably high, but this assumes a pretty sophisticated and risk-averse camp counselor (who, let's be honest, is probably a teenager themselves with an under-developed prefrontal cortex and therefore *highly unikely* to base their decision on a mathematical and risk-averse assessment of the underlying probabilities). Personally, I can't even pretend to understand teenage decision-making and the idea that the counselor's actions would have any relationship to discrete calculations of probabilities is laughable. Who assigns these questions anyway?