DSSH 6301 - HW 04 Solutions

Here are suggested solutions to the homework assignment. Note that many problems have multiple solutions, so we present here merely one example. If you did not have points taken off for a question, we considered your solution correct.

Problem 1

Part a

```
score <- 45
mu <- 70
sigma <- 10

z <- (score - mu) / sigma
z
```

```
## [1] -2.5
```

Part b

```
perc <- pnorm(score, mu, sigma)
perc
```

```
## [1] 0.006209665
```

Part c

```
pnorm(score, mu, sigma) * 2
```

```
## [1] 0.01241933
```

Problem 2

Part a

```
set.seed(101) # Can be any number, it doesn't matter as long as it stays the same.
lambda <- 10
population <- rpois(10000, lambda) # Poisson distribution, but it could be anything.

n <- 9
pop_sample <- sample(population, n, replace=T)
pop_sample
```

```
## [1] 16 12  6  7  9 11 12  8 10
```

Part b

$$n = 9$$

$$total = 16 + 12 + 6 + 7 + 9 + 11 + 12 + 8 + 10 = 91$$

$$\bar{x} = \frac{91}{9} = 10.11111$$

Part c

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s^2 = \frac{1}{8}\sum_{i=1}^{9}(x_i - 10.11111)^2$$

$$s^2 = 9.361111$$

$$s = 3.059593$$

Part d

$$se = \frac{s}{\sqrt{n}}$$

$$se = \frac{3.059593}{\sqrt{9}} = \frac{3.059593}{3} = 1.019864$$

Part e

```
qnorm(c(0.025, 0.975))
```

```
## [1] -1.959964  1.959964
```

$$CI_{0.95} = [\bar{x} - 1.959964 * se, \bar{x} + 1.959964 * se]$$

$$CI_{0.95} = [8.112214, 12.110009]$$

Part e

```
qt(c(0.025, 0.975), n-1)
```

```
## [1] -2.306004  2.306004
```

$$CI_{0.95} = [\bar{x} - 2.306004 * se, \bar{x} + 2.306004 * se]$$

$$CI_{0.95} = [7.7593, 12.46292]$$

Problem 3

Part a

For smaller samples, the mean follows a t distribution rather than a normal distribution. The t distribution in general has thicker tails (higher odds for more extreme values), which results in wider confidence intervals than the normal. In general, one should always use the t, although in practice for $n > 30$ it is close to the normal, and for $n > 100$, almost identical.

Part b

The two errors a person can make here is choosing the wrong number of degrees of freedom and choosing the wrong percentile. The number of degrees of freedom is the sample size n minus 1. The percentile depends on the CI that you are trying to find. For the 90% CI, you need to find the values assocuated with the 5th and 95th percentiles. Since this distribution is symmetric, you can use the value associated with the 5th percentile, $t_{0.05}$.

Problem 4

Part a

If $CI = \bar{x} \pm d$, we want to reduce that to half its size, or $CI_{new} = \bar{x} \pm d_{new} = \bar{x} \pm d/2$.

Since our t statistic doesn't change, that means $se_{new} = se_{old}/2$. Thus

$$\frac{s}{\sqrt{n_{new}}} = \frac{s}{\sqrt{n}} * 1/2$$

Therefore

$$n_{new} = 4 * n$$

So on this case you need a sample size of 36 (4 * 9), 27 more people than the original sample size.

$$se_{new} = \frac{s}{\sqrt{9 * 4}} = \frac{s}{\sqrt{9}} * \frac{1}{2} = se_{old}/2$$

The t-distribution CI:

$$CI_{0.95} = [7.7593, 12.46292]$$

```
old_interval_width <- abs(diff(c(12.46292, 7.7593)))
old_interval_width
```

```
## [1] 4.70362
```

Calculating the new CI:

$$se = \frac{3.059593}{\sqrt{9 * 4}} = 0.5099322$$

t-distribution:

$$CI_{0.95} = [\bar{x} - 2.306004 * se, \bar{x} + 2.306004 * se]$$
$$CI_{0.95} = [8.935205, 11.287017]$$

```
new_interval_width <- abs(diff(c(11.287017, 8.935205)))
new_interval_width
```

```
## [1] 2.351812
```

```
new_interval_width / old_interval_width
```

```
## [1] 0.5000004
```

This of course also holds for the normal distribution. The new CI with a sample size of $4n$ would be

$$CI_{0.95} = [\bar{x} - 1.959964 * se, \bar{x} + 1.959964 * se]$$
$$CI_{0.95} = [9.111662, 11.110560]$$

Part b

We want $\bar{x} \pm d$ such that $d = 1000$. We know that $d = tstat * se$, and $se = s/\sqrt{n}$, so $n = (tstat * se/d)^2$. We can create a simple function to calculate n given any input value of d and s (using the normal approximation of the t here because n is large):

```
num <- function(s, d) (qnorm(0.975)*s / d)^2

s <- 20000

# d = 1000
ceiling(num(s, 1000))
```

```
## [1] 1537
```

```
# d = 100
ceiling(num(s, 100))
```

```
## [1] 153659
```

Note the huge increase in n needed – once again, reducing the interval by a factor of x requires increasing n by a factor of x^2.

Problem 5

```r
nreplicates <- 1000
sample_sz <- 20

sample_summary <- matrix(NA, nrow=nreplicates, ncol=3)

for(j in 1:nreplicates){
  samples <- rexp(sample_sz, 1/lambda) # Exponential population distribution, but it could be anything.

  xbar <- mean(samples)
  se <- sd(samples) / sqrt(sample_sz)
  sample_summary[j,1] <- xbar
  sample_summary[j, 2:3] <- xbar + qt(c(0.005, 0.995), sample_sz-1)*se
}

sum(sample_summary[, 2]<lambda & sample_summary[, 3]>lambda) / nreplicates
```

```
## [1] 0.97
```