BIOS 662

Homework 3 Solution

September, 2018

Question 1

(a) A QQ plot of serum calcium is given on page 3. Most of the points lie close to a straight line but the point in the top right corner is way off the line. Recall that in part 2(c) of Homework 2 it was suggested that one of the calcium values was recorded incorrectly. It is the anomalous value that is in the top right corner. If this value is changed from 18 to 7.5 we get the second of the QQ plots on page 3. In that one all the points lie reasonably close to the line. Later in the semester we will look at a more formal test of normality.

For the rest of the question we will use the uncorrected calcium value.

(b) The population variance is unknown, so we have to handle this either as a small sample from the normal distribution (with unknown variance) or a "large" sample from an unknown distribution. Here $n = 21$, $\bar{Y} = 11.27$, $s^2 = 4.164$ and $s = 2.041$.

If this is a small sample from the normal distribution, a 95% CI for μ is:

$$\bar{Y} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 11.27 \pm 2.086 \times 2.041/\sqrt{21} = (10.34,\ 12.20).$$

If this is a large sample from an arbitrary distribution, using the Central Limit Theorem and Slutsky's Theorem a 95% CI for μ is:

$$\bar{Y} \pm z_{1-\alpha/2}(s/\sqrt{n}) = 11.27 \pm 1.96 \times 2.041/\sqrt{21} = (10.40,\ 12.14).$$

The sample is a little too small to qualify as a large sample and if we leave the value of 18 uncorrected, the normality assumption is also questionable, so neither CI is very satisfactory here.

(c) If in (b) we assume the sample is from the normal distribution, the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.086 \times 2.041/\sqrt{21} = 1.86.$$

Doubling the sample size changes the degrees of freedom of t, so $t_{n-1,1-\alpha/2} = t_{41,0.975} = 2.02$ and the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.02 \times 2.041/\sqrt{42} = 1.27.$$

This is a reduction of $100 \times (1.86 - 1.27)/1.86 = 31.5\%$.

Tripling the sample size, $t_{n-1,1-\alpha/2} = t_{62,0.975} = 2.00$ and the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.00 \times 2.041/\sqrt{63} = 1.03.$$

This is a reduction of $100 \times (1.86 - 1.03)/1.86 = 44.7\%$.

If in (b) we assume the sample is from an arbitrary distribution, the width of the confidence interval is

$$2 \times z_{1-\alpha/2}(s/\sqrt{n}) = 2 \times 1.96 \times 2.041/\sqrt{21} = 1.75.$$

Now doubling the sample size changes just the $\sqrt{n}$ part and the width of the confidence interval becomes

$$2 \times z_{1-\alpha/2}(s/\sqrt{n}) = 2 \times 1.96 \times 2.041/\sqrt{42} = 1.23.$$

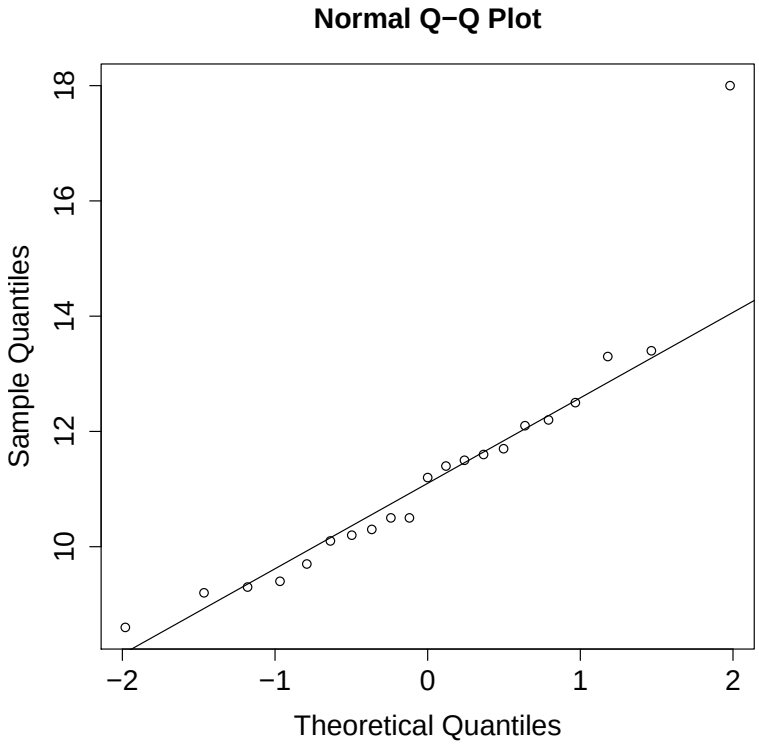
This is a reduction of $100 \times (1.75 - 1.23)/1.75 = 29.3\%$.

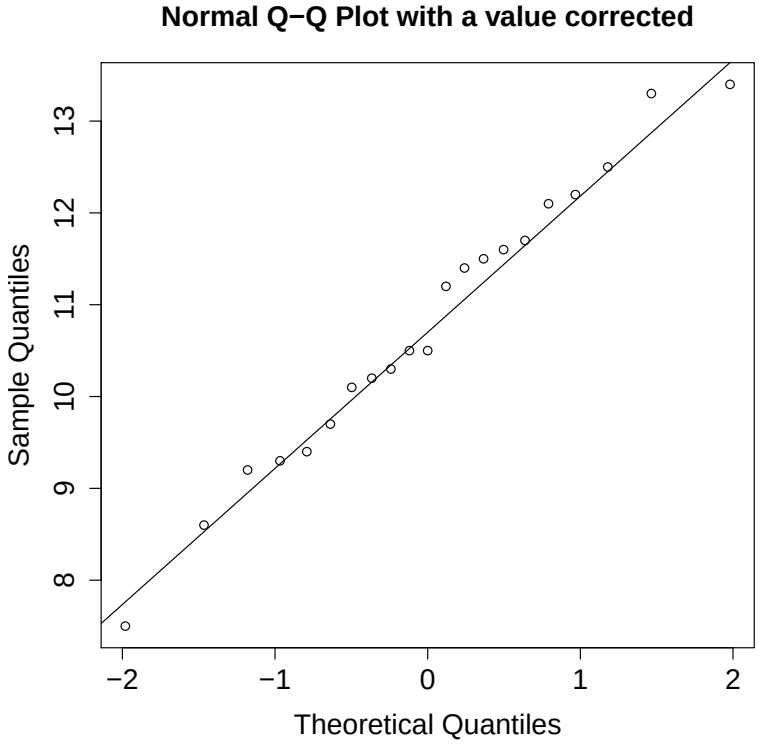
Tripling the sample size the width is $2 \times 1.96 \times 2.041/\sqrt{63} = 1.01$, which is a reduction of $100 \times (1.75 - 1.01)/1.75 = 42.3\%$.

In summary, although the term $t_{n-1,0.975}$ decreases somewhat with increasing sample size, the change in it is relatively small. Most of the change is because of the $1/\sqrt{n}$ part and even then the change is at the rate of $\sqrt{n}$.

(d) The R code I used is given on a subsequent page. (R does have a package, called "boot", for doing boostrapping. But I wanted to demonstrate the principles involved rather than just getting a confidence interval.)

For the particular value in set.seed this yielded 95% CI (10.48, 12.49). Different seeds will yield somewhat different confidence intervals. I tried several values and all gave intervals of the form with lower end 10.4X or 10.5X and upper end 12.4X).


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
10
12
14
16
18
−2
−1
0
1
2


Normal Q−Q Plot with a value corrected
Sample Quantiles
Theoretical Quantiles
8
9
10
11
12
13
−2
−1
0
1
2

R code for the bootstrap-t interval for 1(c):

```
# bootstrap-t interval
set.seed(64353)
mean.Ca<-mean(pge$Ca)
var.Ca<-var(pge$Ca)
n.Ca<-length(pge$Ca)
boots <- 500
zs <- matrix(0,1,boots)
for (jj in 1:boots){
   ysamp <- sample(pge$Ca,size=n.Ca,replace=T)
   zs[jj] <- (mean(ysamp)-mean.Ca)/sqrt(var(ysamp)/n.Ca)
   }
lower.t <- quantile(zs,.975)
upper.t <- quantile(zs,.025)

lower.y <- mean.Ca - lower.t*sqrt(var.Ca/n.Ca)
upper.y <- mean.Ca - upper.t*sqrt(var.Ca/n.Ca)

lower.y
upper.y
```

Programming bootstrap-t intervals in SAS is somewhat more complicated than in R. Below is SAS code, assuming that the data have been read in to a dataset called "pge". For the particular seed used in "proc surveyselect", this yielded 95% CI (10.58, 12.46).

```
proc means data=pge noprint;
  var Ca;
  output out=original mean=original_mean std=original_std n=n;

data original;
  set original;
one=1;  *** For later merging with bootstrap samples;
keep one original_mean original_std n;

proc surveyselect data=pge out=pge_samples seed=45921
    rep=500 sampsize=21 method=urs outhits;
*** rep - specifies the number of replicates
*** method=urs - "requests unrestricted random sampling, which is
***          selection with equal probability and with replacement."
*** outhits - when an observation is selected more than once, the
***          output dataset has a separate row for each occurrence,
***          rather than a single row plus a count of the number of
***          occurrences. ;
```

```
proc means data=pge_samples noprint;
   var Ca;
   by Replicate;
   output out=bootout mean=mean stderr=stderr;

data bootout;
   set bootout;
one=1;

data bootout;
   merge original bootout;
   by one;
zb=(mean-original_mean)/stderr;

proc univariate noprint;
   var zb;
   output out=outpctl pctlpre=P_ pctlpts= 2.5 97.5;

data outpctl;
   set outpctl;
one=1;

data outpctl;
   merge original outpctl;
   by one;

lower=original_mean - P_97_5*original_std/sqrt(n);
upper=original_mean - P_2_5*original_std/sqrt(n);

proc print data=outpctl;
  var lower upper;
```

(e) One way is to do the calculations "manually", using the method on page 43 of the notes on "Point and Interval Estimation". There are 21 patients in the dataset. So we need to find the largest r such that

$$\frac{1}{2^{21}} \sum_{i=0}^{r-1} \binom{21}{i} \leq \alpha/2$$

We can use R to find the largest k such that `sum(dbinom(0:k,21,0.5))` ≤ 0.025 or, equivalently, such that `pbinom(k,21,0.5)` ≤ 0.025.

Because `sum(dbinom(0:5,21,0.5)) = 0.0133` and `sum(dbinom(0:6,21,0.5)) = 0.0392`, $k = 5$ and thus $r - 1 = k = 5$. Hence $r = 6$ and $n - r + 1 = 21 - 6 + 1 = 16$.

A 95% CI for the median is $(X_{(6)}, X_{(16)})$. Sorting the serum calcium values, the 6th and 16th order statistics are 10.1 and 12.1, so the CI is $(10.1, 12.1)$.

Another way is to use SAS:

```
proc univariate cipctldf(type=symmetric);
   var ca;
```

Quantiles (Definition 5)

		95% Conf. Limits		------Order Statistics-----		
Quantile	Estimate	Distribution Free		LCL Rank	UCL Rank	Coverage
100% Max	18.0					
99%	18.0	.	.	.	.	.
95%	13.4	13.3	18.0	19	21	57.45
90%	13.3	12.2	18.0	17	21	83.84
75% Q3	12.1	11.4	13.4	12	20	96.03
50% Median	11.2	10.1	12.1	6	16	97.34
25% Q1	10.1	9.2	10.5	2	10	96.03
10%	9.3	8.6	9.7	1	5	83.84
5%	9.2	8.6	9.3	1	3	57.45
1%	8.6	.	.	.	.	.
0% Min	8.6					

This also has the 95% CI for the median as $(10.1, 12.1)$. The 97.34 in the "Coverage" column is obtained as $100 \cdot (1 - 2 \times 0.0133) = 97.34$.

Question 2 – Problem 4.20 on page 111

(a) If $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then $\bar{Y} \sim N(\mu, \sigma^2/n)$.

Here $\mu = 1.0$, $\sigma^2 = 9.0$ and $n = 9$, so $\bar{Y} \sim N(1, 9/9) = N(1,1)$. That is, the sampling distribution of $\bar{Y}$ is normal with mean 1 and variance 1.

(b) Standardizing by subtracting the mean of $\bar{Y}$ and dividing by the standard error,

$$\begin{aligned} \Pr[1 < \bar{Y} \leq 2.85] &= \Pr\left[\frac{1-1}{\sqrt{1}} < \frac{\bar{Y}-1}{\sqrt{1}} \leq \frac{2.85-1}{\sqrt{1}}\right] \\ &= \Pr[0 < Z \leq 1.85] = \Pr[Z \leq 1.85] - \Pr[Z \leq 1] \\ &= \Phi(1.85) - \Phi(1) = 0.9678 - 0.5 = 0.4678. \end{aligned}$$

(c) Using properties on page 22 of the notes on "Statistical Inference: Populations and Samples", if $\bar{Y} \sim N(1,1)$ and $W = 4\bar{Y}$, then $W \sim N(4 \times 1, 4^2 \times 1) = N(4, 16)$.