

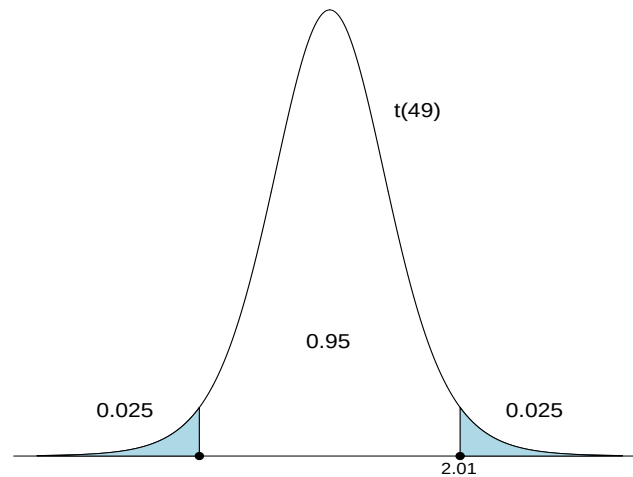
1. (a) A 95% confidence interval for the population mean μ is

$$\bar{x} \pm t_{49,0.025} \frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and $n = 50$ is the sample size. The quantile

$$t_{49,0.025} \approx 2.01$$

is the upper 2.5th percentile (i.e., 97.5th percentile) of the $t(49)$ distribution:



```
> options(digits=3)
> qt(0.975,49)
[1] 2.01
```

We have

$$\bar{x} \pm t_{49,0.025} \frac{s}{\sqrt{n}} \implies 685.7 \pm 2.01 \left(\frac{168.8}{\sqrt{50}} \right) \implies (637.7, 733.7).$$

Interpretation: We are 95% confident the population mean lifetime μ is between 637.7 and 733.7 hours. Based on this analysis, I would recommend against the purchase. The confidence interval consists entirely of values less than 750 hours. The advertiser's claim of "having an average lifetime of 750 hours" is not supported by this analysis.

(b) There are two assumptions associated with the confidence interval in part (a):

1. The sample of 50 light bulbs is a random sample from the population of all light bulbs (so that the sample is "representative" of the population)
2. The lifetime X for each bulb is normally distributed (i.e., the population distribution is normal).

The normal qq plot shown in the assignment looks like there is good agreement between the data and a normal distribution (the qq plot is largely linear). Of course, the confidence interval in part (a) is robust to departures from normality, so even if there were slight deviations, this wouldn't cause large problems with our analysis or conclusions.

(c) We will first find a 95% confidence interval for the population variance σ^2 and then take the square root of each endpoint of the interval. A 95% confidence interval for the population variance σ^2 is

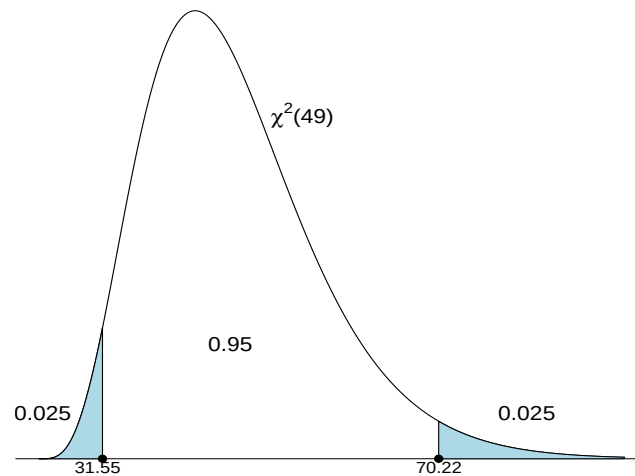
$$\left(\frac{(n-1)s^2}{\chi_{49,0.975}^2}, \frac{(n-1)s^2}{\chi_{49,0.025}^2} \right),$$

where s^2 is the sample variance and $n = 50$ is the sample size. The quantiles

$$\chi_{49,0.025}^2 \approx 31.55$$

$$\chi_{49,0.975}^2 \approx 70.22$$

are the 2.5th and 97.5th percentiles of the $\chi^2(49)$ distribution:



```
> qchisq(0.025,49)
[1] 31.55
> qchisq(0.975,49)
[1] 70.22
```

We have

$$\left(\frac{(n-1)s^2}{\chi_{49,0.975}^2}, \frac{(n-1)s^2}{\chi_{49,0.025}^2} \right) \Rightarrow \left(\frac{49(168.8)^2}{70.22}, \frac{49(168.8)^2}{31.55} \right) \Rightarrow (19882.9, 44252.9).$$

We are 95% confident the population variance σ^2 is between 19882.9 and 44252.9 (hours)². A 95% confidence interval for the population standard deviation σ is

$$(\sqrt{19882.9}, \sqrt{44252.9}) \Rightarrow (141.0, 210.4).$$

Interpretation: We are 95% confident the population standard deviation σ is between 141.0 and 210.4 hours.

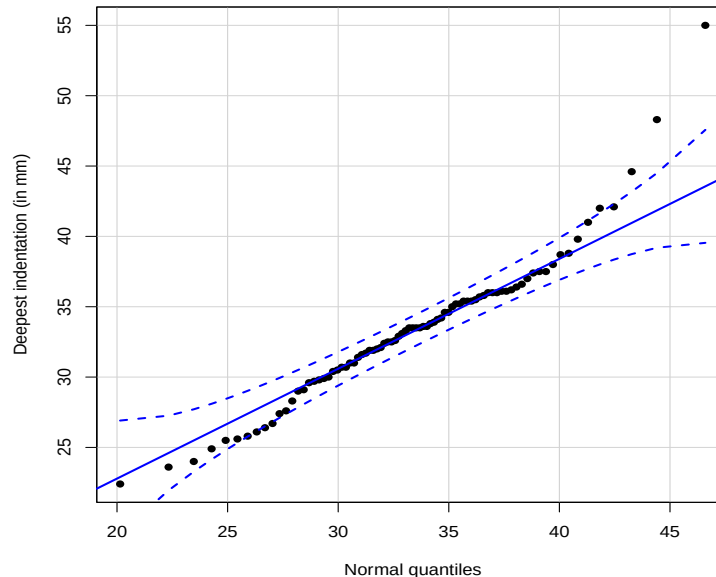
The assumptions for this interval are the same as they were for the confidence interval for the population mean μ in part (a). The only difference is that confidence intervals for σ^2 and σ depend critically on the normal distribution assumption for the population of all light bulbs. Thankfully, the normal qq plot in the assignment doesn't show any cause for concern here.

2. (a) I used the `t.test` function in R to calculate a 95% confidence interval for the population mean μ :

```
> t.test(indent, conf.level=0.95)$conf.int
[1] 32.2 34.5
```

Interpretation: We are 95% confident the population mean deepest indentation μ is between 32.2 and 34.5 mm.

Here is a normal qq plot for the data:



Assessment: This plot shows a general agreement between the observed data and the normal quantiles. There are a couple of “larger than expected” observations in the upper tail, but this isn’t too concerning. The confidence interval above is robust to mild and even moderate normality departures anyway.

(b) Your co-worker is misinterpreting the confidence interval! He’s right that only 17 out of the 83 observations are between 32.2 and 34.5 mm. However, a confidence interval has nothing to do with this. This confidence interval is an interval where we infer the population mean μ resides—not where the individual measurements are.

(c) We will first find a 95% confidence interval for the population variance σ^2 and then take the square root of each endpoint of the interval. I will use the `var.ci` function I wrote in R to find a 95% confidence interval for the population variance σ^2 .

```
> options(digits=3)
> var.ci(indent, conf.level=0.95) # CI for population variance
[1] 20.9 38.7
```

Interpretation: We are 95% confident the population variance σ^2 is between 20.9 and 38.7 $(\text{mm})^2$.

```
> options(digits=2)
> sqrt(var.ci(indent, conf.level=0.95)) # CI for population standard deviation
[1] 4.6 6.2
```

Interpretation: We are 95% confident the population standard deviation σ is between 4.6 and 6.2 mm.

Confidence intervals for σ^2 and σ strictly require the normal population distribution assumption (and are not robust to violations of this assumption). One might be a little concerned about these intervals being accurate given the minor normality departures we saw earlier in the qq plot. However, the departure we saw was so slight (e.g., just a couple of outliers on the high side) that I wouldn't be that worried.

(d) This part is asking you to determine the sample size n (for a larger study) to estimate the population mean indentation μ with these requirements:

- 99% confidence $\implies z_{0.005} \approx 2.58$.

```
> options(digits=3)
> qnorm(0.995,0,1)
[1] 2.58
```

- margin of error $B = 1.15/2 = 0.575$.

– Our confidence interval in part (a) has length 2.3 mm. One-half of this is 1.15 mm. The margin of error B is half the interval length.

- What should we use as a “guess” of the population standard deviation σ ? Here it would be clearly sensible to use the sample standard deviation s as an estimate of what σ is:

```
> options(digits=2)
> sd(indent)
[1] 5.3
```

You could also pick values of σ which reside in the 95% confidence interval in part (c). All of these values of σ are consistent with the sample information.

For the values specified above, we would use

$$n = \left(\frac{z_{\alpha/2}\sigma}{B} \right)^2 = \left(\frac{2.58 \times 5.3}{0.575} \right)^2 \approx 565.5.$$

The larger study would require $n = 566$ clay models to meet the requirements outlined above.

3. (a) It's hard to know exactly which type of women the CDC and state health departments are targeting. A reasonable answer is all women in the US who have recently given birth.

(b) The sample proportion of women who smoked during the last 3 months of pregnancy is

$$\hat{p} = \frac{31}{250} \approx 0.124.$$

A 95% confidence interval for the population proportion p is

$$0.124 \pm 1.96 \sqrt{\frac{0.124(1-0.124)}{250}} \implies 0.124 \pm 0.041 \implies (0.083, 0.165).$$

Among all women in the US who have recently given birth, we are 95% confident the proportion who smoked during the last 3 months of pregnancy is between 0.083 and 0.165.

(c) This part is asking you to determine the sample size n (for a larger study) to estimate the population proportion p with these requirements:

- 99% confidence $\implies z_{0.005} \approx 2.58$.
- margin of error $B = 0.02$.
- What should we use as a “guess” of the population proportion p_0 ? Here it would be clearly sensible to use the sample proportion $\hat{p} = 0.124$ as a guess. You could also pick various values of p_0 which reside in the 95% confidence interval in part (b). All of these values are consistent with the sample information. You could be completely conservative and choose $p_0 = 0.5$, but this seems unnecessary here.

For the selections above, we have

$$n = \left(\frac{z_{\alpha/2}}{B} \right)^2 p_0(1 - p_0) = \left(\frac{2.58}{0.02} \right)^2 (0.124)(1 - 0.124) \approx 1807.6.$$

The larger study would require $n = 1807$ women to meet the requirements outlined above.

ADDITIONAL R CODE:

```
# Problem 1(a)
# t(49) with quantiles
x = seq(-5,5,0.001)
pdf = dt(x,10)
plot(x,pdf,type="l",lty=1,xlab="",xaxt="n",yaxt="n",bty="n",ylab="",ylim=c(0,0.4))
abline(h=0)
x = seq(-5,qt(0.025,10),0.001)
y = dt(x,10)
polygon(c(-5,x,qt(0.025,10)),c(0,y,0),col="lightblue")
points(x=qt(0.025,10),y=0,pch=19,cex=1)
x = seq(qt(0.975,10),5,0.001)
y = dt(x,10)
polygon(c(qt(0.975,10),x,5),c(0,y,0),col="lightblue")
points(x=qt(0.975,10),y=0,pch=19,cex=1)
text(-0.025,0.1,0.95,cex=1.25)
text(-3.5,0.04,0.025,cex=1.25)
text(3.5,0.04,0.025,cex=1.25)
text(1.5,0.3,"t(49)",cex=1.25)
text(2.21,-0.011,2.01,cex=1)
```

```
# Problem 1(c)
# chi^2(49) with quantiles
x = seq(0,30,0.001)
pdf = dchisq(x,10)
plot(x,pdf,type="l",lty=1,xlab="",ylab="",xaxt="n",yaxt="n",bty="n",ylim=c(0,0.1))
abline(h=0)
```

```
x = seq(0,qchisq(0.025,10),0.001)
y = dchisq(x,10)
polygon(c(0,x,qchisq(0.025,10)),c(0,y,0),col="lightblue")
points(x=qchisq(0.025,10),y=0,pch=19,cex=1)
x = seq(qchisq(0.975,10),30,0.001)
y = dchisq(x,10)
polygon(c(qchisq(0.975,10),x,30),c(0,y,0),col="lightblue")
points(x=qchisq(0.975,10),y=0,pch=19,cex=1)
text(9.8,0.025,0.95,cex=1.25)
text(0.18,0.01,0.025,cex=1.25)
text(25,0.01,0.025,cex=1.25)
text(13.5,0.075,expression(paste(chi^2, "(49)")),cex=1.25)
text(3,-0.002,31.55,cex=1)
text(20.5,-0.002,70.22,cex=1)

# Problem 2
indent = c(22.4,23.6,24.0,24.9,25.5,25.6,25.8,26.1,26.4,26.7,27.4,27.6, 28.3,
29.0, 29.1, 29.6, 29.7, 29.8, 29.9, 30.0, 30.4, 30.5, 30.7, 30.7, 31.0, 31.0,
31.4, 31.6, 31.7, 31.9, 31.9, 32.0, 32.1, 32.4, 32.5, 32.5, 32.6, 32.9, 33.1,
33.3, 33.5, 33.5, 33.5, 33.5, 33.6, 33.6, 33.8, 33.9, 34.1, 34.2, 34.6, 34.6,
35.0, 35.2, 35.2, 35.4, 35.4, 35.4, 35.5, 35.7, 35.8, 36.0, 36.0, 36.0, 36.1,
36.1, 36.2, 36.4, 36.6, 37.0, 37.4, 37.5, 37.5, 38.0, 38.7, 38.8, 39.8, 41.0,
42.0, 42.1, 44.6, 48.3, 55.0)

# Problem 2(a)
t.test(indent,conf.level=0.95)$conf.int
library(car)
qqPlot(indent,distribution="norm",mean=mean(indent),sd=sd(indent),
        xlab="Normal quantiles",ylab="Deepest indentation (in mm)",pch=16,
        envelope=list(border=TRUE,style="lines"),id=FALSE)

# Problem 2(c)
# CI for population variance function
var.ci = function(data,conf.level=0.95){
  df = length(data)-1
  chi.lower = qchisq((1-conf.level)/2,df)
  chi.upper = qchisq((1+conf.level)/2,df)
  s2 = var(data)
  c(df*s2/chi.upper,df*s2/chi.lower)
}
options(digits=3)
var.ci(indent,conf.level=0.95) # CI for population variance
options(digits=2)
sqrt(var.ci(indent,conf.level=0.95)) # CI for population standard deviation
```