1. (a) From the boxplots, we can see the samples look to have different amounts of variation. The boxplot for Location 1 looks to be rather "wide," whereas the Location 2 boxplot looks much more compact. This is not what we would expect to see if the population variances were equal, that is, $\sigma_1^2 = \sigma_2^2$. Therefore, I would opt for the confidence interval which does not assume equal population variances.

We could construct a confidence interval for the population variance ratio

$$\Lambda = \frac{\sigma_2^2}{\sigma_1^2}.$$

This interval is based on the F distribution. We could then look to see if "1" is included in this interval. If the interval excludes "1," this is good evidence the population variances are not equal.
(b) I chose the unequal-variance confidence interval for $\Delta = \mu_1 - \mu_2$, the population mean difference.

Interpretation: We are 95% confident the difference of the population mean number of organisms (per m^2) between the two locations is between 1389 and 10164. Because this interval for $\Delta = \mu_1 - \mu_2$ contains only positive values, this is consistent with the population mean of Location 1 (μ_1) being larger than the population mean of Location 2 (μ_2).

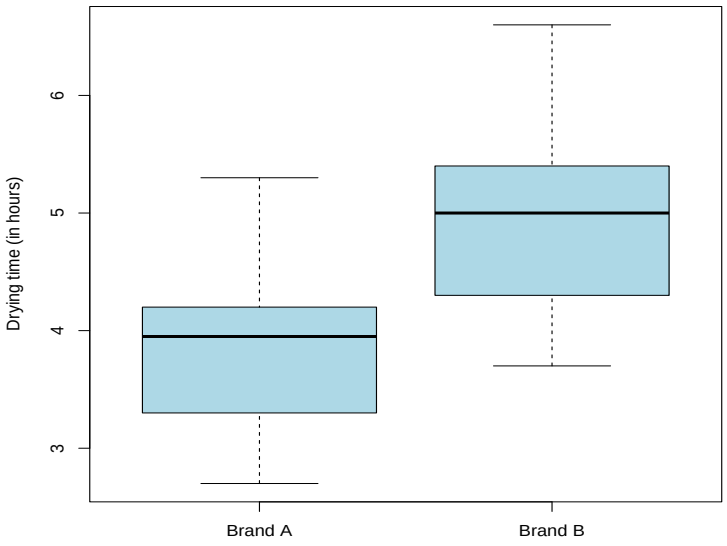
(c) The boxplots look like distributions which are skewed to the right (high) side. You can see this in the very long upper tail in Location 1 and the off-center location of the median in the Location 2. The qq-plots are probably picking this up. Remember the normal distribution is symmetric−any skewness in the samples will produce disagreement in the qq-plots.

Even though there may be moderate departures from normality (in one sample or both), remember confidence intervals for population means and population mean differences are generally robust to these departures. This is a consequence of the CLT. This means we can still use confidence intervals for population means even though the underlying normality assumption may not hold exactly.

Again, as noted in the assignment, trying to make definite "yes-no" decisions on what the population distributions are using small samples (like 12 and 16) is a pretty tall order. If we had larger samples, we could make a more informed assessment.

2. The first thing to do is look at the samples graphically. I used side-by-side boxplots to do this; see the top of the next page. Here are some initial observations:

- The boxplots look pretty symmetric in shape (no apparent skewness to one side or the other). This bodes well for the underlying normality assumption for both populations. We will check this later using qq plots.
- The variation in the boxplots looks roughly the same for each sample. This bodes well for using the equal-variance confidence interval for the population mean difference $\Delta = \mu_1 - \mu_2$. Remember the goal of this analysis is to see how the population mean drying times compare.



We can go ahead and formally look at how the population variances compare. We can do this by writing a confidence interval for the population variance ratio

$$\Lambda = \frac{\sigma_2^2}{\sigma_1^2}.$$

I used R's `var.test` function to write a 95% confidence interval for Λ using the two samples:

```
> options(digits=2)
> var.test(Brand.B,Brand.A,conf.level=0.95)$conf.int
[1] 0.38 2.22
```

We are 95% confident the population variance ratio Λ is between 0.38 and 2.22. Not surprisingly, this interval includes "1," which would correspond to equal population variances.

I used R's `t.test` function to write a 95% confidence interval for $\Delta = \mu_1 - \mu_2$, the difference of the population means (1 = Brand A; 2 = Brand B):

```
> t.test(Brand.A,Brand.B,conf.level=0.95,var.equal=TRUE)$conf.int
[1] -1.59 -0.64
```

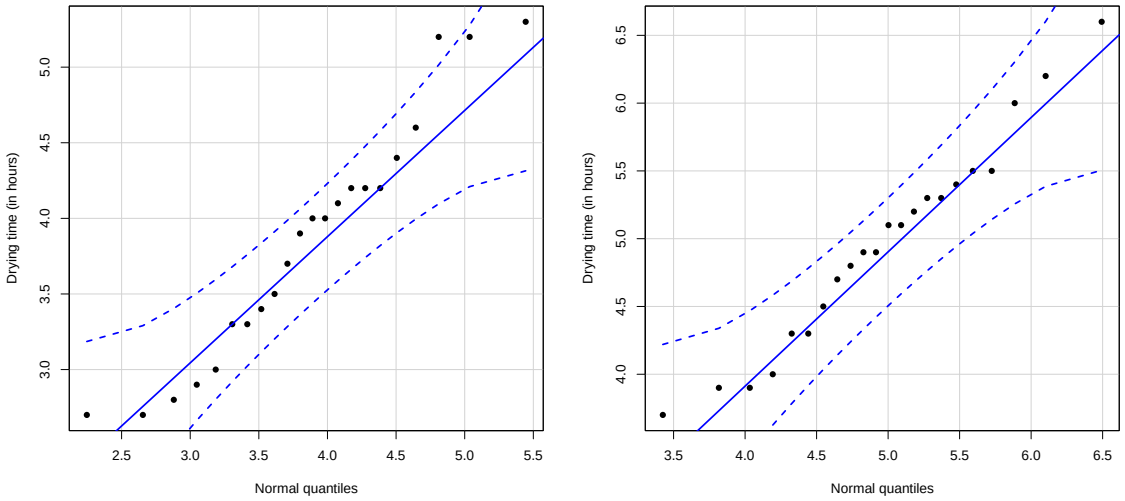
Interpretation: We are 95% confident the difference of the population mean drying times is between -1.59 and -0.64 hours. Because this interval for $\Delta = \mu_1 - \mu_2$ contains only negative values, this is consistent with the population mean drying time of Brand A (μ_1) being smaller than the population mean drying time of Brand B (μ_2).

Note: Interestingly, the 95% confidence interval for $\Delta = \mu_1 - \mu_2$ which does not assume the population variances are equal gives the same interval (to 2 dp):

```
> t.test(Brand.A,Brand.B,conf.level=0.95,var.equal=FALSE)$conf.int
[1] -1.59 -0.64
```

Here are the assumptions that we are making in this analysis:

- The two samples are random samples. This means the 44 pieces of pressure-treated wood were randomly selected from a larger population and the paint used is representative of Brand A and Brand B paint.
- Independent samples. This is a reasonable assumption because the pieces of wood were randomized to receive either Brand A or Brand B paint.
- The population distribution of drying time is normal for both brands. We can diagnose this assumption by using qq-plots:



We see general agreement between the observed data (Brand A = left; Brand B = right) and the normal quantiles. I don't have any concerns about normality violations.

Summary: A two-group analysis was performed to compare the population mean drying time of two types of paint (Brand A and Brand B). The analysis shows strong evidence the population mean drying time for Brand A paint is less than the population mean drying time for Brand B paint (95% CI for difference: $-1.59 < \Delta < -0.64$ hours). All of the statistical assumptions which this analysis requires appear to be satisfied.

3. (a) Point estimates for p_1 and p_2 based on this study are

$$\widehat{p}_1 = \frac{8}{86} \approx 0.093 \quad \text{and} \quad \widehat{p}_2 = \frac{10}{142} \approx 0.070,$$

and a 95% confidence interval for $\Delta = p_1 - p_2$ is

$$(0.093 - 0.070) \pm 1.96\sqrt{\frac{0.093(1-0.093)}{86} + \frac{0.070(1-0.070)}{142}} \longrightarrow (-0.051, 0.097).$$

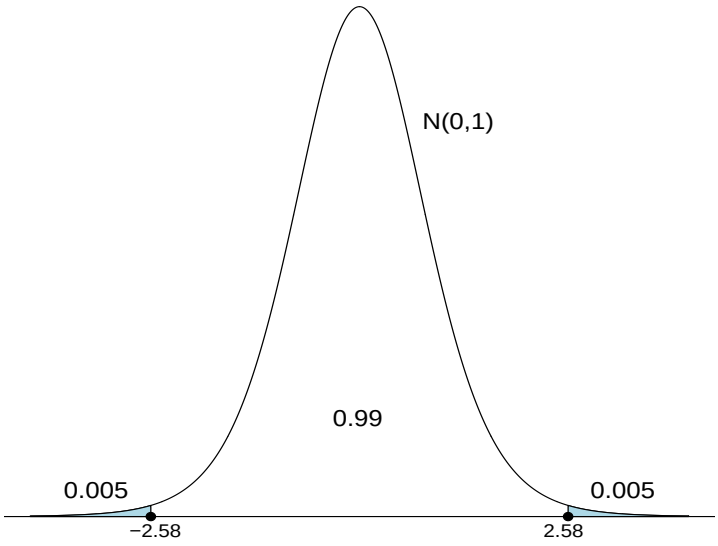
Interpretation: We are 95% confident the population proportion difference $\Delta = p_1 - p_2$ is between -0.051 and 0.097. Because this interval contains "0," we do not have sufficient evidence (at the 95% confidence level) that the population proportion of exceedences is different for the two airlines.

Note that we can also use the `prop.test` function in R to calculate this confidence interval directly:

```
> options(digits=3)
> > prop.test(c(8,10),c(86,142),conf.level=0.95,correct=FALSE)$conf.int
[1] -0.052  0.097
```

It looks like a small amount of rounding error was introduced in my hand calculations above.

(b) Let's first determine the value $z_{\alpha/2}$ which corresponds to 99% confidence:




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

We now set the margin of error for the 99% confidence interval

$$z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}=0.03$$

and solve for n. We have $z_{\alpha/2}\approx 2.58$ and the estimates $\widehat{p}_1$ and $\widehat{p}_2$ from part (a). Thus,

$$\begin{aligned}
2.58\sqrt{\frac{0.093(1-0.093)}{n}+\frac{0.070(1-0.070)}{n}}=0.03 \quad &\Longrightarrow \quad \sqrt{\frac{0.149}{n}}=\frac{0.03}{2.58} \\
&\Longrightarrow \quad \frac{0.149}{n}=\left(\frac{0.03}{2.58}\right)^2 \\
&\Longrightarrow \quad \frac{n}{0.149}=\left(\frac{2.58}{0.03}\right)^2 \\
&\Longrightarrow \quad n=0.149\left(\frac{2.58}{0.03}\right)^2\approx 1102.004.
\end{aligned}$$

We would need to observe about 1102 landings for each airline to write a 99% confidence interval for the population proportion difference $\Delta=p_1-p_2$ with margin of error equal to 0.03.

ADDITIONAL R CODE:

```
# Problem 2
# Enter the data
Brand.A = c(3.5,2.7,3.9,4.2,4.6,2.7,3.3,5.2,4.2,2.9,4.4,
    5.2,4.0,4.1,3.4,3.3,4.2,5.3,3.7,3.0,4.0,2.8)
Brand.B = c(4.7,3.9,4.5,5.5,4.0,5.3,4.3,6.0,5.3,3.7,5.5,
    6.2,5.1,5.4,4.8,4.9,6.6,4.3,4.9,5.1,3.9,5.2)

# Side-by-side boxplots
boxplot(Brand.A,Brand.B,xlab="",names=c("Brand A","Brand B"),
    ylab="Drying time (in hours)",col="lightblue")

# Confidence interval for population variance ratio
var.test(Brand.B,Brand.A,conf.level=0.95)$conf.int

# Confidence interval for the population mean difference
t.test(Brand.A,Brand.B,conf.level=0.95,var.equal=TRUE)$conf.int

# qq plots for normality
library(car)
qqPlot(Brand.A,distribution="norm",mean=mean(Brand.A),sd=sd(Brand.A),
    xlab="Normal quantiles",ylab="Drying time (in hours)",pch=16,
    envelope=list(border=TRUE,style="lines"),id=FALSE)
qqPlot(Brand.B,distribution="norm",mean=mean(Brand.B),sd=sd(Brand.B),
    xlab="Normal quantiles",ylab="Drying time (in hours)",pch=16,
    envelope=list(border=TRUE,style="lines"),id=FALSE)


# Problem 3(b)
# N(0,1) 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(0.005,10),0.001)
y = dt(x,10)
polygon(c(-5,x,qt(0.005,10)),c(0,y,0),col="lightblue")
points(x=qt(0.005,10),y=0,pch=19,cex=1)
x = seq(qt(0.995,10),5,0.001)
y = dt(x,10)
polygon(c(qt(0.995,10),x,5),c(0,y,0),col="lightblue")
points(x=qt(0.995,10),y=0,pch=19,cex=1)
text(-0.025,0.075,0.99,cex=1.25)
text(-4,0.02,0.005,cex=1.25)
text(4,0.02,0.005,cex=1.25)
text(1.5,0.3,"N(0,1)",cex=1.25)
text(3.1,-0.011,2.58,cex=1)
text(-3.1,-0.011,-2.58,cex=1)
```