DATA 602 - Solutions to Assignment Two

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(mosaic)
```

```
## Loading required package: mosaic
```

```
## Warning: package 'mosaic' was built under R version 3.4.4
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: lattice
```

```
## Loading required package: ggformula
```

```
## Warning: package 'ggformula' was built under R version 3.4.4
```

```
## Loading required package: ggstance
```

```
## Warning: package 'ggstance' was built under R version 3.4.4
```

```
##
## Attaching package: 'ggstance'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
## Loading required package: mosaicData
```

```
## Warning: package 'mosaicData' was built under R version 3.4.4
```

```
## Loading required package: Matrix
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affec
ted by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median,
##     prop.test, quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
require(binom)
```

```
## Loading required package: binom
```

1. (From Question 11, Assignment 1) **(2 marks)**. From Assignment 1, the delivery time is modeled by the Normal distribution with a mean of $\mu = 5.0$ hours and a standard deviation of $\sigma = 1.5$ hours. A random sample of $n = 12$ produced a sample mean of $\overline{X} = 5.6875$.

a. **Answer** Here one wishes to find $P(\overline{X} \geq 5.6875)$, where the distribution of $\overline{X}$ *will be exactly* Normal with

$$\mu_{\overline{X}} = \mu_X = 5.0 \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{1.5}{\sqrt{12}} = 0.4330 \approx 0.433$$

$P(\overline{X} \leq 5.6875)$ is then computed

```
1 - pnorm(5.6875, 5.0, (1.5/sqrt(12)))
```

```
## [1] 0.0561756
```

and

$$P(\overline{X} \geq 5.6875) = 0.0562$$

- **1 mark** for the correct answer. Jiang, if the student does not divide by $\sigma_{\overline{X}}$, award zero marks.

b. **Answer:**

$$
\begin{aligned}
P(0.5 \leq S \leq 1) =& P(0.5^2 \leq S^2 \leq 1^2) \\
=& P\left(\frac{(n-1) * 0.5^2}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1) * 1^2}{\sigma^2} \right) \\
=& P\left(\frac{(12-1) * 0.5^2}{1.5^2} \leq \chi^2_{df=12-1} \leq \frac{(12-1) * 1^2}{1.5^2} \right) \\
=& P(1.222 \leq \chi^2_{df=12-1} \leq 4.889) \\
=& 0.0634
\end{aligned}
$$

This probability is computed in R

```
pchisq(4.889, 11) - pchisq(1.222, 11)
```

```
## [1] 0.06343895
```

- **1 mark** for the correct answer.

---

2. (5 marks)

a. **Answer**. The mean and standard deviation of the distribution of $\widehat{p}$ is

$$
\mu_{\widehat{p}} = p = 0.80 \qquad \sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{500}} = \sqrt{\frac{0.80(1-0.80)}{500}} = 0.0179
$$

- **(1 mark)**, 0.5 mark for each of the mean and standard deviation of the sample proportion.

b. **Answer** Compute $P(\widehat{p} \leq 0.748)$ via R Studio

```
pnorm(0.748, 0.80, 0.0179)
```

```
## [1] 0.001836102
```

and $P(\widehat{p} \leq 0.748) = 0.001836 \approx 0.0018$

- **(1 mark)** (Jiang, if the student computed $P(\widehat{p} \leq 0.744)$, then award marks, There was a typo in the original posting of this assignment.

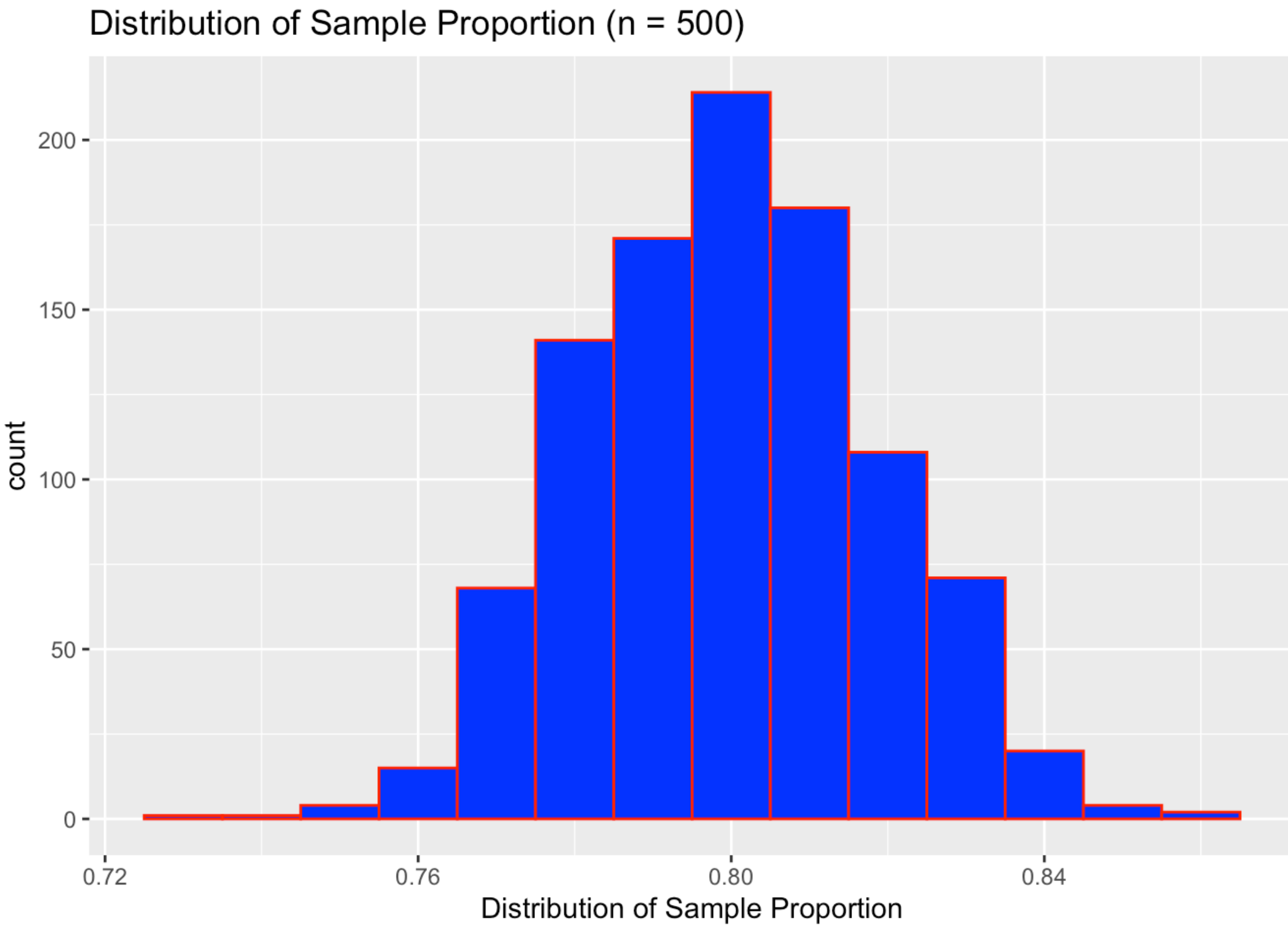
c. **Answer:**

```
ntimes = 1000
ntrials = 500
propsupport = numeric(ntimes)
propobserved = numeric(ntimes)
for(i in 1:ntimes)
{  propsupport[i] = (rbinom(1, ntrials, 0.80)/ntrials)
   if (propsupport[i] <= 0.744) propobserved[i] = 1 else propobserved[i] = 0
   }
ass2q2 = data.frame(propsupport, propobserved)
head(ass2q2, 4)
```

	propsupport <dbl>	propobserved <dbl>
1	0.788	0
2	0.814	0
3	0.776	0
4	0.832	0

4 rows

```
ggplot(data=ass2q2, aes(x = propsupport)) + geom_histogram(fill='blue', col='red', bi
nwidth=0.01) + xlab("Distribution of Sample Proportion") + ggtitle("Distribution of S
ample Proportion (n = 500)")
```



(2 marks) for the generation of the distribution of the sample proportion (which should roughly be the same as above)

```
# proportion of sample proportions that are less than 0.748
sum(~ propobserved, data=ass2q2)/ntimes  #OR
```

```
## [1] 0.002
```

```
sum((propsupport <= 0.748))/ntimes
```

```
## [1] 0.003
```

(1 mark) for computing the proportion of sample proportions that are less than the observed value of $\widehat{p} = 0.748$. (Jiang, results will differ from one student to the next, but they should be in neighbourhood of 0.001 - 0.003)

3. (3 marks)

Answer Solutions should be in the following structure.

One has to consider Billy's claim, that $\overline{X} > 1$. To compute "how likely" Billy's claim is, we invoke the Central Limit Theorem, where the distribution of the mean number of matching numbers $\overline{X}$ is approximately Normally distributed with a mean and standard deviation of

$$\mu_{\overline{X}} = \mu_X = 0.7347 \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{0.7599}{\sqrt{52}} = 0.105379 \approx 0.1054$$

To compute the probability of Billy's claim:

```
options(scipen=999)
1 - pnorm(1, 0.7347, 0.1054)
```

```
## [1] 0.005916635
```

and $P(\overline{X} > 1) = 0.005917 \approx 0.0059$, which is very unlikely.

- **(2 marks)** for finding the probabilty of Billy's claim

Therefore, Billy's claim is not support from a probability perspective.

- **(1 mark)** for a comment on Billy's claim, the comment being supported from the probability computation

---

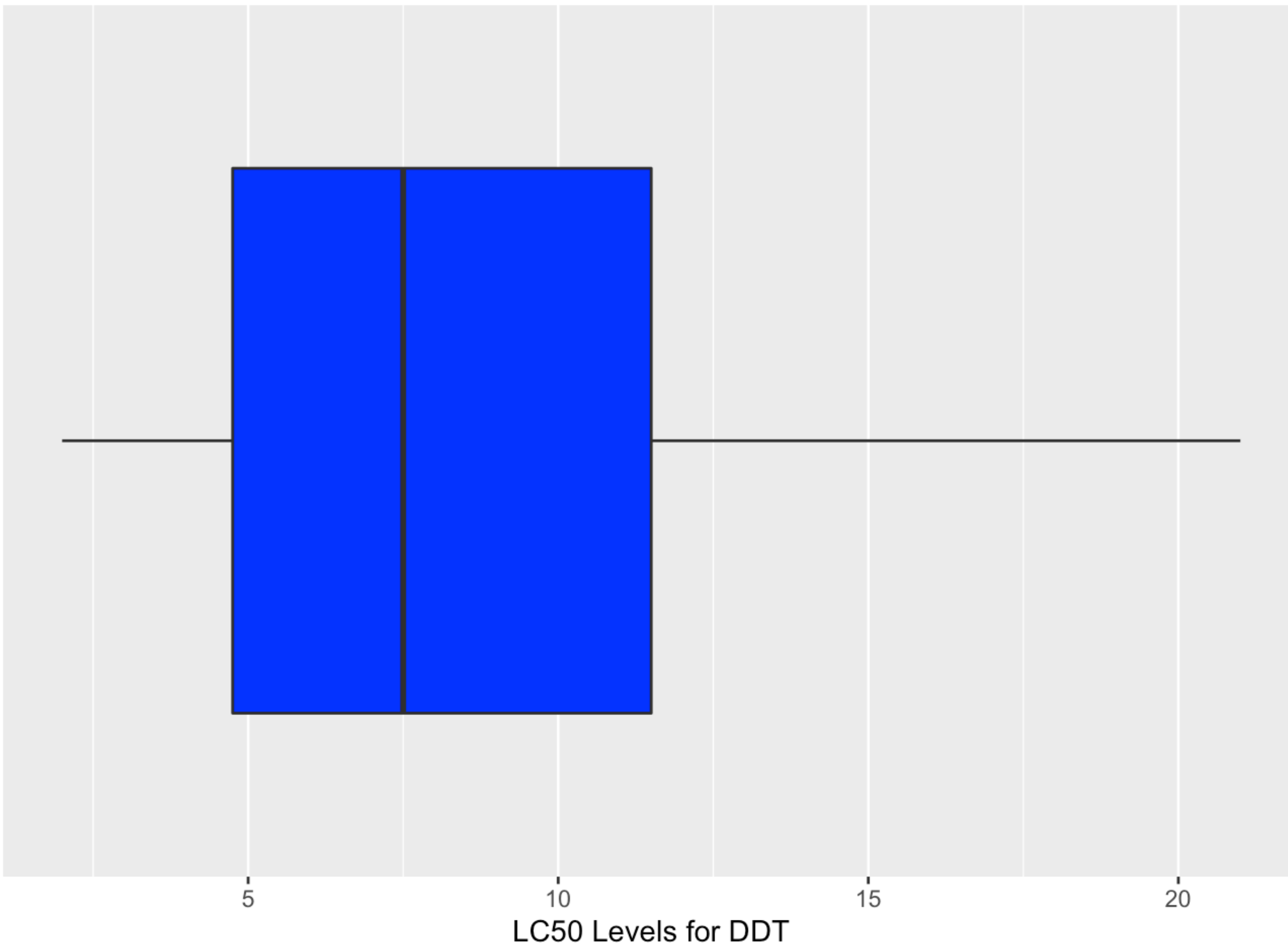
4. (10 marks)

```
lc50 = c(16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9)
ass2q4df = data.frame(lc50)
head(ass2q4df, 4)
```

	lc50 <dbl>
1	16
2	5
3	21
4	19

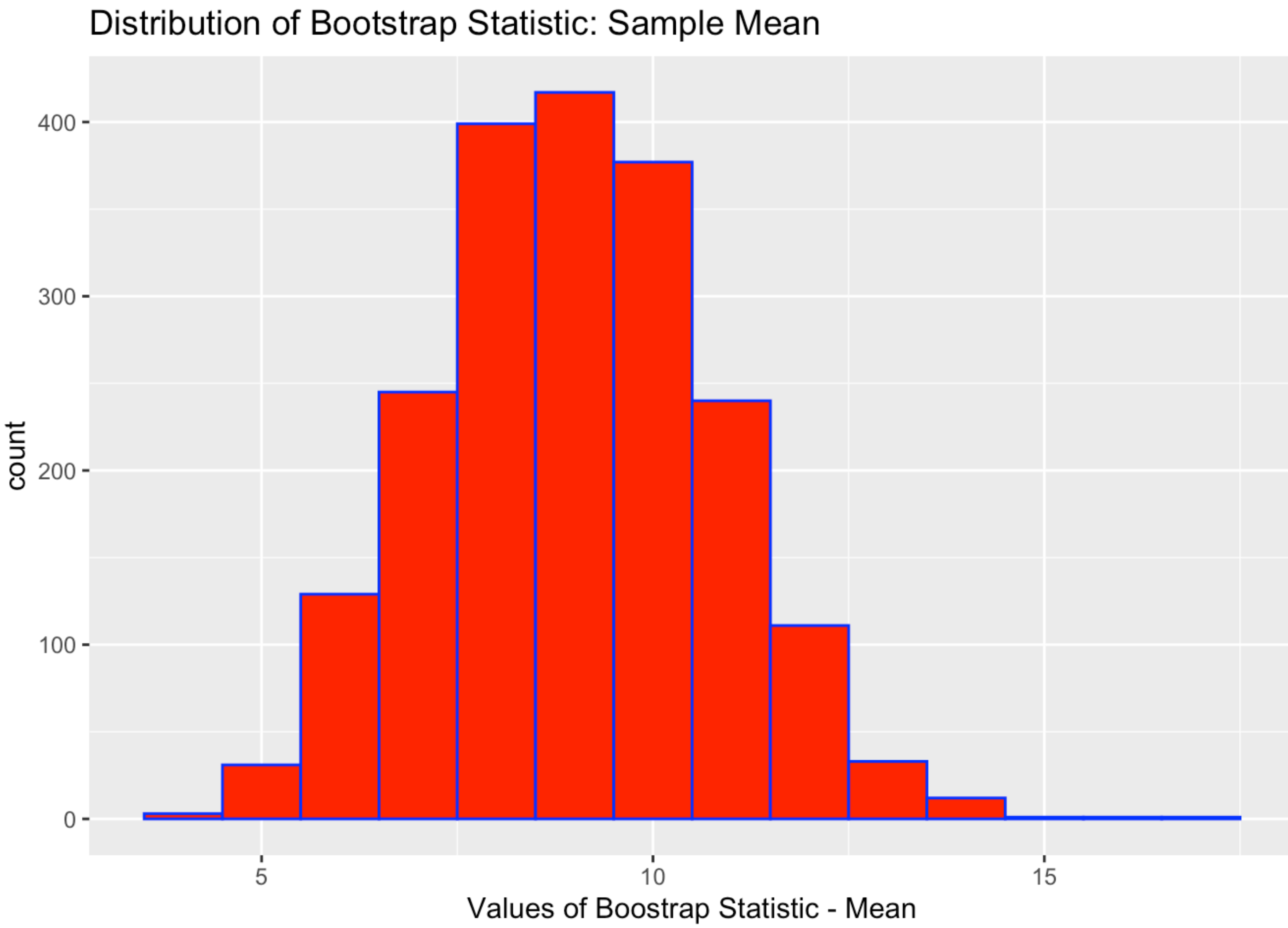
4 rows

```
ggplot(ass2q4df) + geom_boxplot(mapping = aes(x = "var", y =lc50), fill= 'blue', na.r
m=TRUE) + xlab("") + ylab("LC50 Levels for DDT") + scale_x_discrete(breaks=NULL) + co
ord_flip()
```



a. **Answer:**

```
nsims = 2000
ntrials = 12
avelc50 = numeric(nsims)
for(i in 1:nsims)
{    avelc50[i] = mean(sample(lc50, ntrials, replace=TRUE))
}
ass2q4 = data.frame(avelc50)
ggplot(data = ass2q4, aes(x = avelc50)) + geom_histogram(fill='red', col='blue', binw
idth=1) + xlab("Values of Boostrap Statistic - Mean") + ggtitle("Distribution of Boot
strap Statistic: Sample Mean")
```



- **(2 marks)** (Jiang, use your judgement here, as the results will vary from one student to the next. The distribution should be close to symmetrical with a central value around 9. If the student has roughly the same result award **2 marks**; any moderate deviations penalize **1 mark**)

b. **Answer:**

```
qdata(~ avelc50, c(0.025, 0.975), data=ass2q4)
```

	quantile <dbl>	p <dbl>
2.5%	5.75	0.025
97.5%	12.50	0.975

2 rows

The 95% bootstrap interval for μ, the mean amount of DDT required to kill 50% of the certain species of fish within 96 hours of exposure is somewhere beween 5.667 ppm and 12.585 ppm.

- **(1 mark)** for providing a 95% interval from their bootstrap distribution. As long as the student outlines/used either the **qdata()** or **quantile()** command to obtain the 2.5th and the 97.5th percentile from their bootstrap distribution in part (a), award full marks here.
- **(1 mark)** for the correct interpretation. Within this interpretation, ensure the student interprets their interval with the condition "from these data/based on these data (0.5 mark)" *AND* the student does indicate that the confidence interval is a narrowing down of the possible values of the population mean μ, that it, the mean is some value between the lower bound and the upper bound.

c. **Answer:** Using the t.test() command

```
t.test(~ lc50, conf.level=0.95, data = ass2q4df)$conf
```

```
## [1]  4.91814 13.08186
## attr(,"conf.level")
## [1] 0.95
```

Using the t-interval, the 95% confidence interval is: $4.9181 \leq \mu \leq 13.082$.

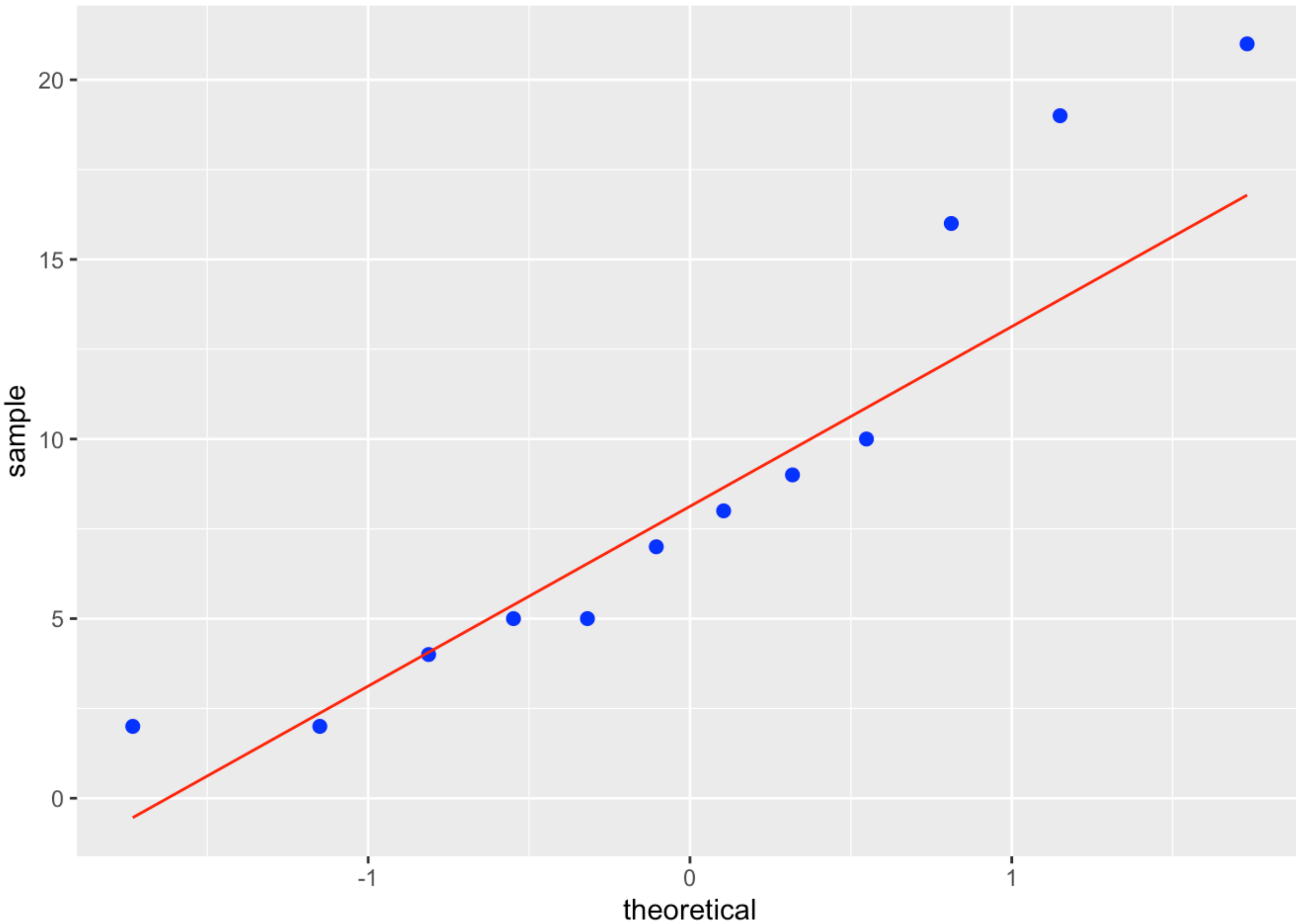
- **(2 marks)**. One mark for the correct value of the lower bound and 1 mark for the correct value of the upper bound.

d. **Answer:**

- **(2 marks)** Jiang, you will get a variety of answers here. The bootstrap interval is "condition free", meaning it does not depend on the condition of Normality of the data and takes into account the "non-perfect" bootstrap distribuition of the sample mean which shows some skewness to the right. As long as the student has commented about this, award full marks.

e. **Answer:**

```
ggplot(data=ass2q4df, aes(sample = lc50)) + stat_qq(size=2, col='blue') + stat_qqline
(col='red')
```



[1^]: http://angusreid.org/wp-content/uploads/2015/02/2015.02.13-Vaccinations.pdf (http://angusreid.org/wp-content/uploads/2015/02/2015.02.13-Vaccinations.pdf)

- **(1 mark)** for the generation of the Normal probability plot
- **(1 mark)** for a comment on the data following a Normal distribution (this will be subjective, it would appear that these data are not Normally distributed from the absence of linearity, but some may view the bulk-middle as being linear, and that is fine).

---

5. (7 marks)

a. **Answer:** A 95% confidence interval for p, using the “plus-2/plus-4” version is computed using the **binom.confint()** command:

```
binom.confint(571, 1866, method="agresti-coull")
```

method	x	n	mean	lower	upper
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

1	agresti-coull	571	1866	0.3060021	0.2855056	0.3272958

1 row

```
#OR
binom.test(571, 1866, ci.method="plus4")
```

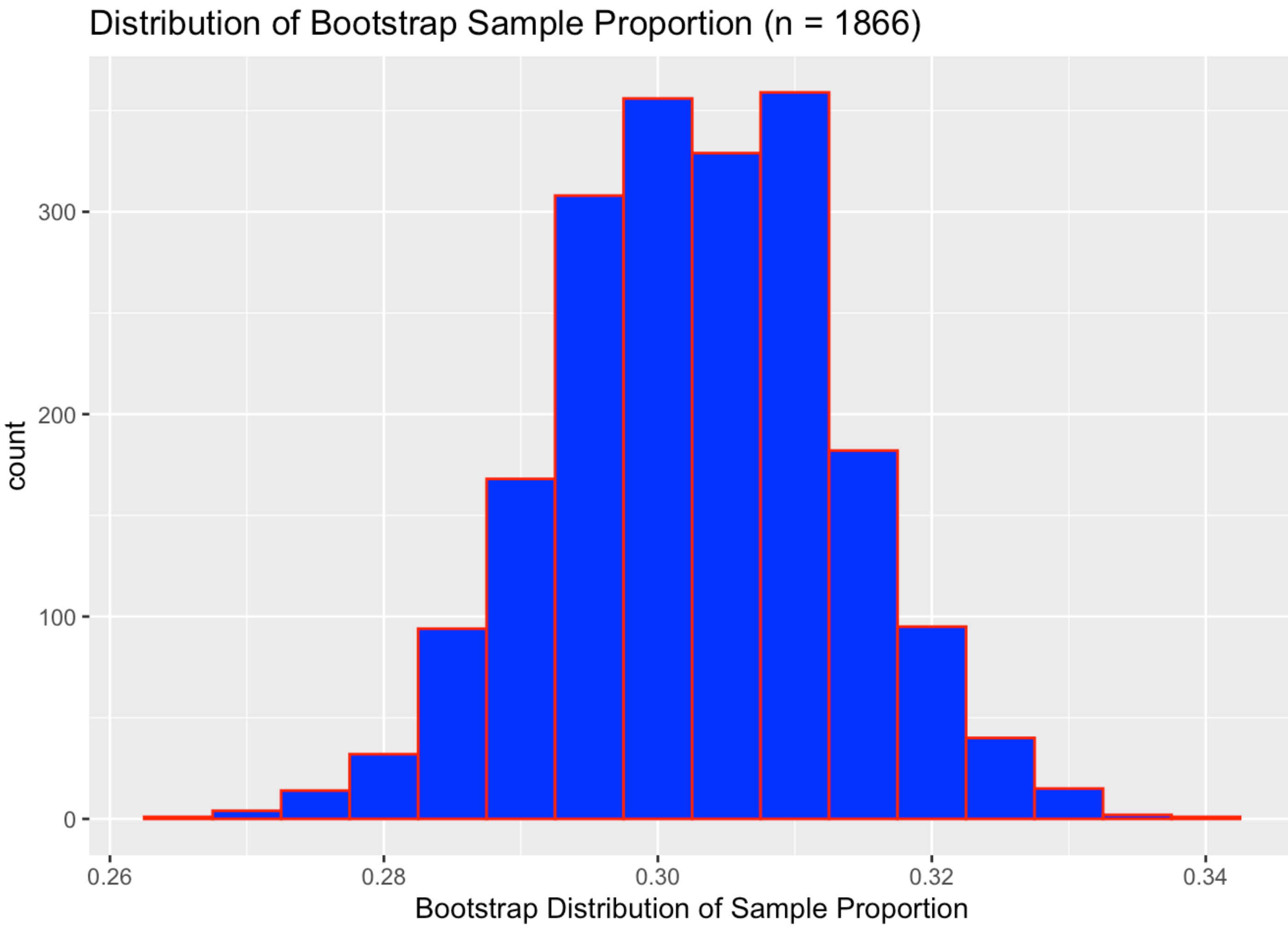
```
##
##  Exact binomial test (Plus 4 CI)
##
## data:  571 out of 1866
## number of successes = 571, number of trials = 1866, p-value <
## 0.00000000000000022
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2855226 0.3273117
## sample estimates:
## probability of success
##              0.3060021
```

The 95% confidence interval for p is $ 0.29 p 0.33$.

- **(2 marks)**. 1 mark for the correct lower bound, 1 mark for the correct upper bound

b. **Answer** Below is the code to create a bootstrap distribution of the sample proportion $\widehat{p}$:

```
nsims = 2000
nsize=1866
sampleprop = numeric(nsims)
ques2data = c(rep(0, 1886 - 571), rep(1, 571)) #create a data vectore of 571 1s and (
1866 - 571) 0s
for(i in 1:nsims)
{   sampleprop[i] = sum(sample(ques2data, nsize, replace=TRUE))/(nsize)
}
ques5df = data.frame(sampleprop)
# head(ques2df, 3)
ggplot(ques5df, aes(x = sampleprop)) + geom_histogram(col='red', fill='blue', binwidt
h=0.005) + xlab("Bootstrap Distribution of Sample Proportion") + ggtitle("Distributio
n of Bootstrap Sample Proportion (n = 1866)")
```



- **(2 marks)** Mark similar to Question 1(a). If the student provides a bootstrap distribution of the sample proportion which should appear to be similar to provided, award full (2) marks here Ajmery.

c. **Answer:** To obtain a 95% bootstrap CI for p, obtain the 2.5 and 97.5 percentiles:

```
qdata(~ sampleprop, c(0.025, 0.975), data=ques5df)
```

	quantile <dbl>	p <dbl>
2.5%	0.2824223	0.025
97.5%	0.3231511	0.975

2 rows

The 95% confidence interval is $0.2819 \leq p \leq 0.3231$.

- **(1 mark)** Jiang, mark similar to Question 4(b). As long as the student obtains their interval from the bootstrap distribuion of $\widheat p$, award 1 mark.

d. **Answer:** The bootstrap interval of $[0.28819, 0.3231]$ compared to the plus2/plus4 of $[0.28, 0.33]$? Both are similar width, the bootstrap distribution showing some skewness to the left would be the preferred interval here.

- **(2 marks)** Jiang, please mark Question 5(d) similar to Question 4(e). As long as the student provides a “statistical justification” for their answer, award 2 marks.

---

6. (9 marks)

a. **Answer:** Below is the bootstrap distribution for the $\widehat{p}_{HS}$.

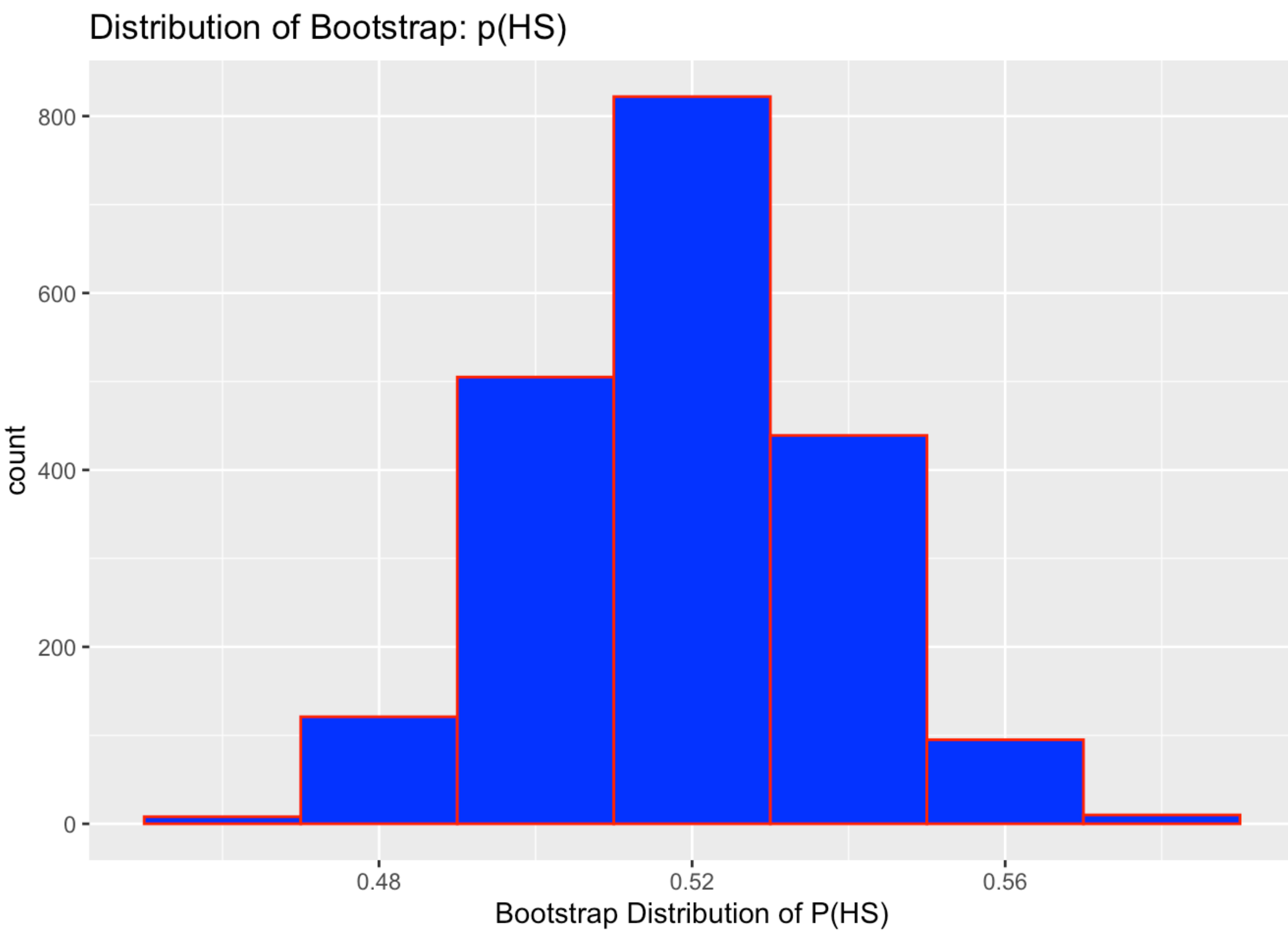
```
nsims = 2000
hsdata = c(rep(0,670-348), rep(1, 348)) #hs data with 348 1s and (670-348) 0s
hssampleprop = numeric(nsims)
for(i in 1:nsims)
{ hssampleprop[i] = sum(sample(hsdata, 670, replace=TRUE))/670
}
ques6adf = data.frame( hssampleprop)
head(ques6adf, 3)
```

	hssampleprop <dbl>
1	0.5701493
2	0.5194030
3	0.5552239

3 rows

Below is a histogram of the bootstrap statistic $\widehat{p}_{HS}$

```
ggplot(data=ques6adf, aes(x = hssampleprop)) + geom_histogram(col='red', fill='blue',
binwidth=0.02) + xlab("Bootstrap Distribution of P(HS)") + ggtitle("Distribution of B
ootstrap: p(HS)")
```



- **(1 mark)** Mark similar to how students were marked in generating the bootstrap distributions in both Question 1 and Question 2.

b. **Answer:** Below is the bootstrap distribution for the $\widehat{p}_{Uni}$.

```
nsims = 2000
unidata = c(rep(0,376-274), rep(1, 274)) #university data with 274 1s and (376-274) 0
s
unisampleprop = numeric(nsims)
for(i in 1:nsims)
{
   unisampleprop[i] = sum(sample(unidata, 376, replace=TRUE))/376
}
ques6bdf = data.frame(unisampleprop)
head(ques6bdf, 3)
```

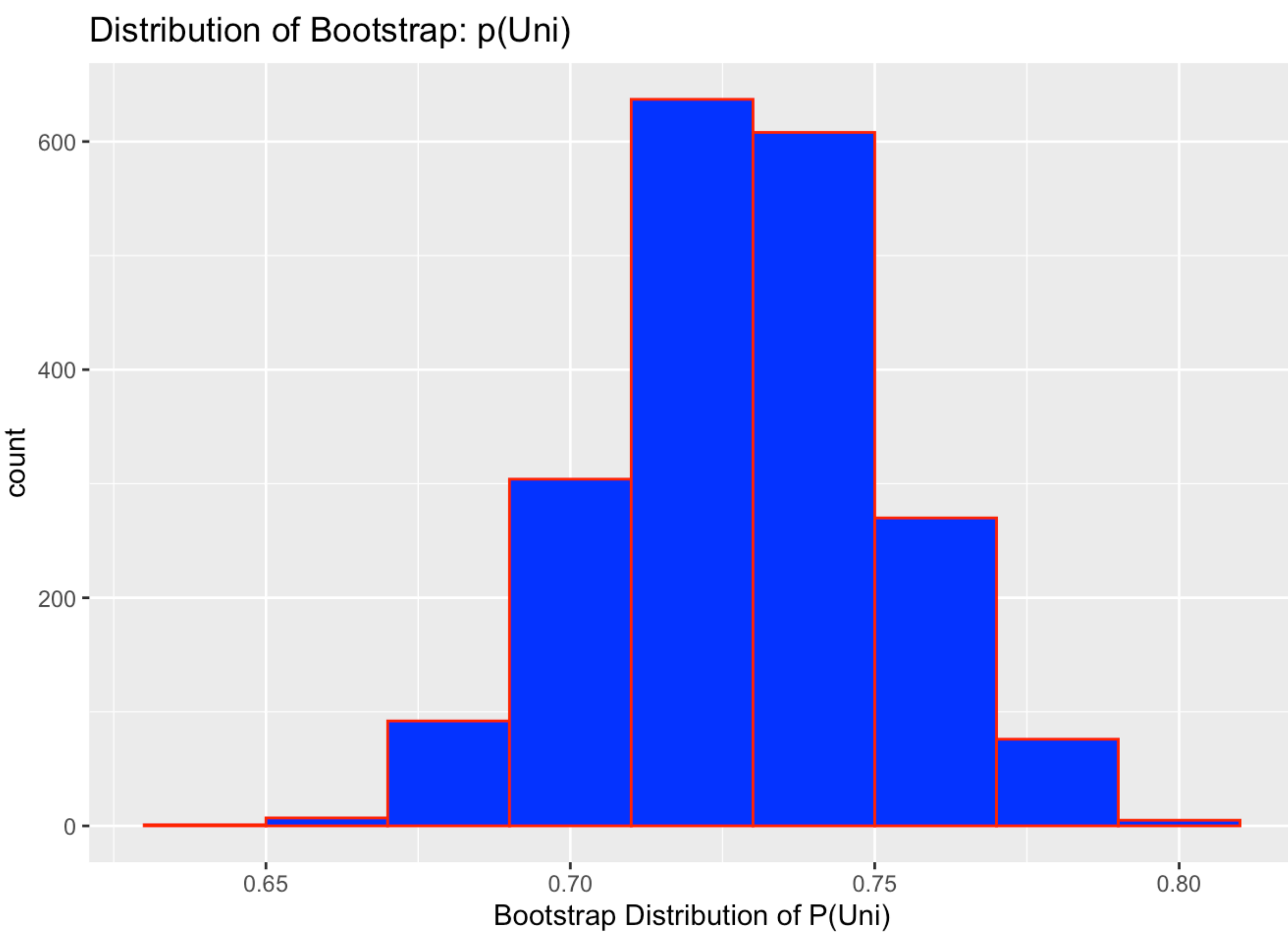
unisampleprop
<dbl>

1	0.7287234
2	0.7393617
3	0.7367021

3 rows

Below is a histogram of the bootstrap statistic $\widehat{p}_{Uni}$

```
ggplot(data=ques6bdf, aes(x = unisampleprop)) + geom_histogram(col='red', fill='blue'
, binwidth=0.02) + xlab("Bootstrap Distribution of P(Uni)") + ggtitle("Distribution o
f Bootstrap: p(Uni)")
```



- **(1 mark)** Mark similar to how students were marked in generating the bootstrap distributions in part (a)

c. **Answer:** Below is the bootstrap distribution for the $\widehat{p}_{Uni} - \widehat{p}_{HS}$.

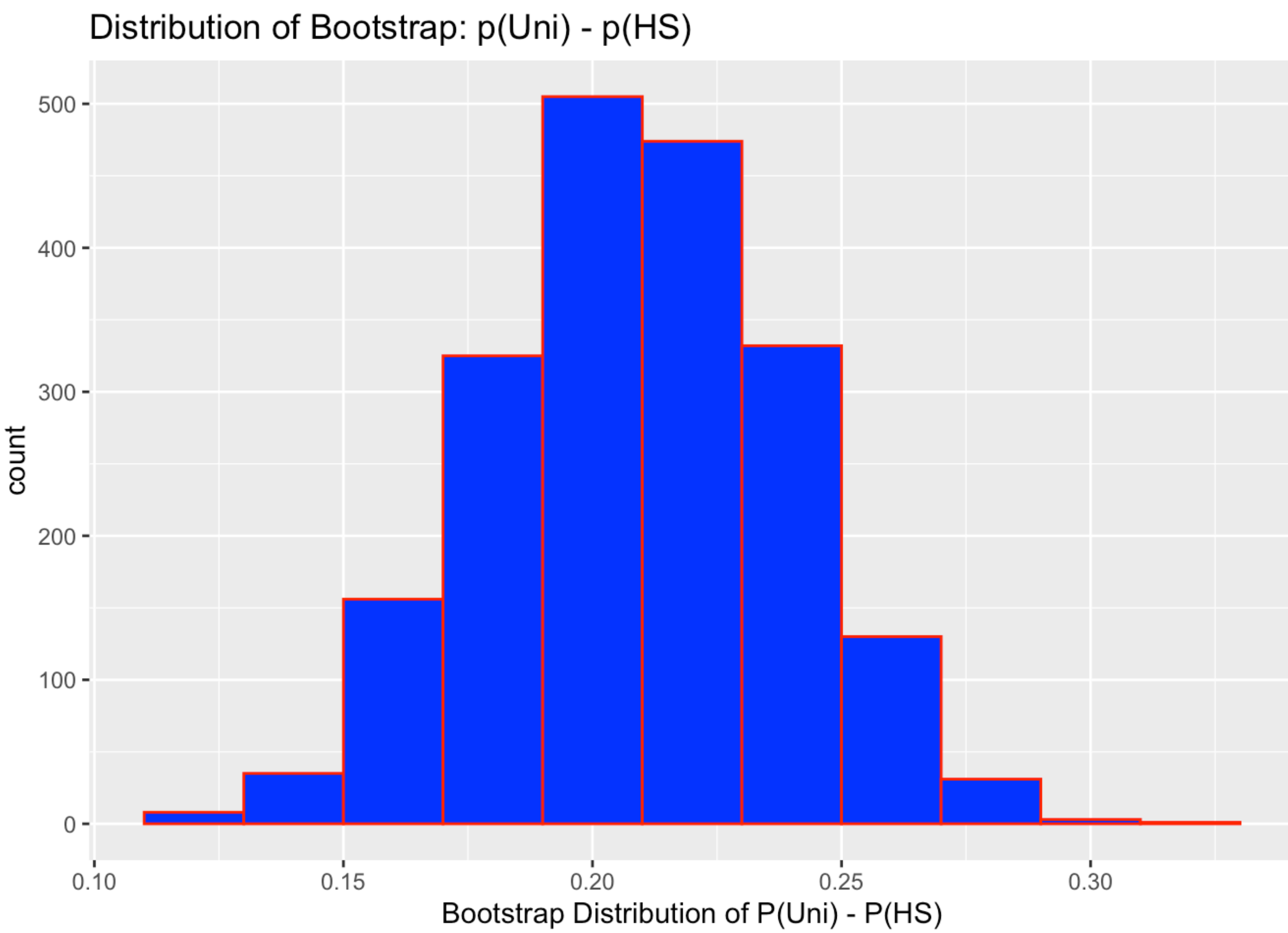
```
nsims = 2000
unidata = c(rep(0,376-274), rep(1, 274)) #university data with 274 1s and (376-274) 0
s
hsdata = c(rep(0,670-348), rep(1, 348)) #hs data with 348 1s and (670-348) 0s
unisampleprop = numeric(nsims)
hssampleprop = numeric(nsims)
diffsampleprop = numeric(nsims)
for(i in 1:nsims)
{
   unisampleprop[i] = sum(sample(unidata, 376, replace=TRUE))/376
   hssampleprop[i] = sum(sample(hsdata, 670, replace=TRUE))/670
   diffsampleprop[i] = unisampleprop[i] - hssampleprop[i]
}
ques6df = data.frame(unisampleprop, hssampleprop, diffsampleprop)
head(ques6df, 3)
```

	unisampleprop <dbl>	hssampleprop <dbl>	diffsampleprop <dbl>
1	0.7393617	0.5358209	0.2035408
2	0.7606383	0.5029851	0.2576532
3	0.7420213	0.4955224	0.2464989

3 rows

Below is a histogram of the bootstrap statistic $\widehat{p}_{Uni} - \widehat{p}_{HS}$

```
ggplot(data=ques6df, aes(x = diffsampleprop)) + geom_histogram(col='red', fill='blue'
, binwidth=0.02) + xlab("Bootstrap Distribution of P(Uni) - P(HS)") + ggtitle("Distri
bution of Bootstrap: p(Uni) - p(HS)")
```



- **(3 marks)** Mark similar to how students were marked in generating the bootstrap distributions in both Question 1 and Question 2.

d. **Answer:**

From this, the 95% bootstrap interval is

```
qdata(~ diffsampleprop, c(0.025, 0.975), data=ques6df)
```

	quantile <dbl>	p <dbl>
2.5%	0.1513814	0.025
97.5%	0.2653872	0.975

2 rows

$$0.1508 \leq p_{Uni} - p_{Hs} \leq 0.2683$$

- **(2 mark)** for the provision of the bootstrap interval, 1 mark for the lower bound, and 1 mark for the upper bound.

No, one cannot infer that $p_{Uni} = p_{HS}$, as the confidence interval has a lower bound of approximately 0.149 and an upper bound of about 0.27. Because this confidence inteval has a lower bound that exceeds 0, one can infer that p_{Uni} **EXCEEDS** p_{HS} by anywhere from 14.9% to 27%.

- **(2 marks)** Jiang, as long as the student makes a statement that the CI does not capure zero, hence one cannot conclude that $p_{Uni} = p_{HS}$, award 2 marks.

---

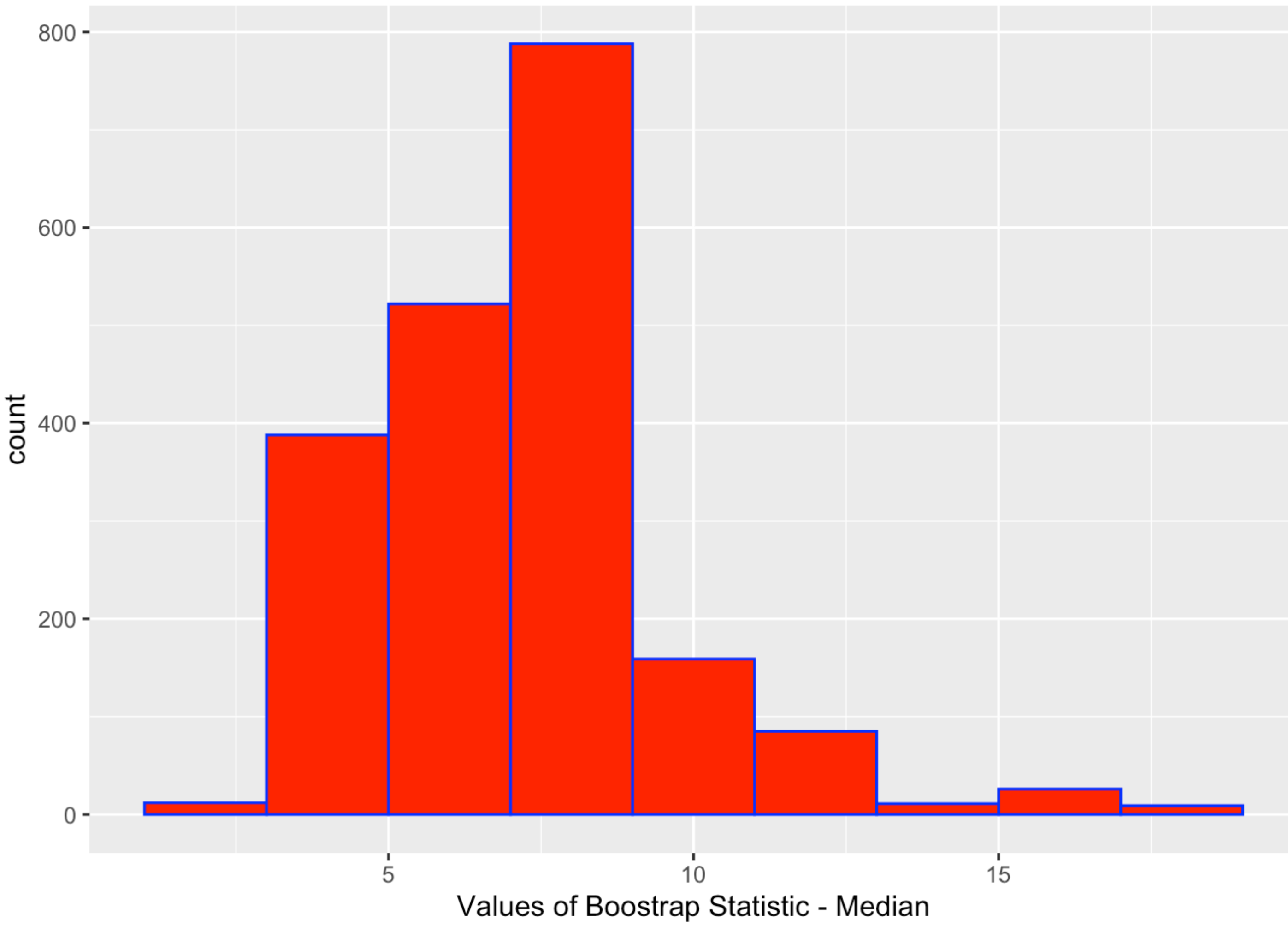
7. (3 marks)

Re-using the code from Question 4, changing the **mean** to **median**

```
lc50 = c(16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9)
nsims = 2000
ntrials = 12
medianlc50 = numeric(nsims)
for(i in 1:nsims)
{    medianlc50[i] = median(sample(lc50, ntrials, replace=TRUE))
}
ass3q4 = data.frame(medianlc50)
```

The bootstrap distribution of the sample median $\widetilde{X}$ appears below.

```
ggplot(data = ass3q4, aes(x = medianlc50)) + geom_histogram(fill='red', col='blue', b
inwidth=2) + xlab("Values of Boostrap Statistic - Median")
```



- **(2 marks)** for creating the boostrap distribution of the sample median.

From this, a 99% confidence interval for $\widehat{\mu}$ is $3.5 \leq \widetilde{\mu} \leq 16$.

```
qdata(~ medianlc50, c(0.005, 0.995), data=ass3q4)
```

	quantile <dbl>	p <dbl>
0.5%	3	0.005
99.5%	16	0.995

2 rows

```
quantile(ass3q4$medianlc50, c(0.005, 0.995))
```

```
##  0.5% 99.5%
##     3    16
```

- **(1 marks)** 0.5 mark for the correct lower bound and 0.5 mark for the correct upper bound *based* on the student's usage of the **qdata()** or the **quantile()** command

---

8. (7 marks)

a. **Answer:** 95% confidence interval for p_{NDP} is

```
binom.test(126, 1003, ci.method="plus4")$conf
```

```
## [1] 0.1065370 0.1476835
## attr(,"conf.level")
## [1] 0.95
## attr(,"method")
## [1] "plus4"
```

and

$$0.1065 \leq p_{NDP} \leq 0.1476$$

- **2 marks**, 1 for the correct lower bound and 1 for the correct upper bound

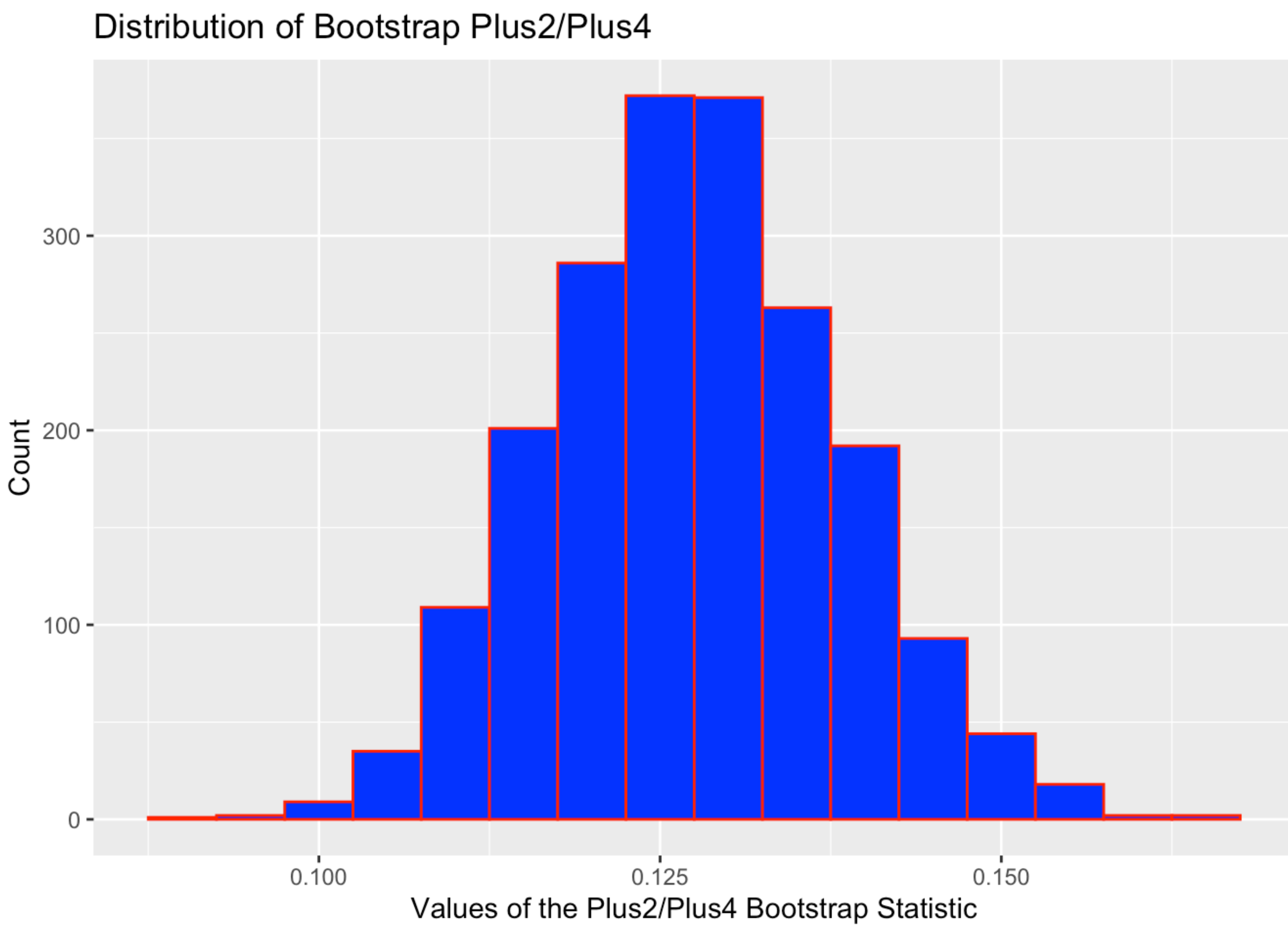
b. **Answer:** The bootstrap distribution of $\frac{X_{NDP}+2}{n+4}$ is provided below

```
Nsims = 2000
nsize = 1003
ndpdata= c(rep(0, 1003 - 126), rep(1, 126)) #puts the data in the form of (1003 - 126
) 0s and 126 1s
```

Here are the contents of the bootstrap:

```
bootptilde = numeric(Nsims)
for(i in 1:Nsims)
{
  bootptilde[i] = (sum(sample(ndpdata, nsize, replace=TRUE)) + 2)/(nsize + 4)
}
bootq8 = data.frame(bootptilde)
```

```
ggplot(data=bootq8, aes(x = bootptilde)) + geom_histogram(col='red', fill='blue', bin
width = 0.005) + xlab("Values of the Plus2/Plus4 Bootstrap Statistic") + ylab("Count"
) + ggtitle("Distribution of Bootstrap Plus2/Plus4")
```



- **2 marks** for generating the bootstrap distribution of $\frac{X_{NDP}+2}{n+4}$

c. **Answer:** The 95% bootstrap interval for p from the result in (b) is

```
qdata(~bootptilde, c(0.025, 0.975), data=bootq8)
```

	quantile <dbl>	p <dbl>
2.5%	0.1082423	0.025
97.5%	0.1489573	0.975

2 rows

the 95% bootstrap interval for p is then

$$0.1072 \leq p \leq 0.1470$$

- **1 mark** for the bootstrap inteval. Again, results will vary from one student to the next.

d. Answers here will vary Jiang. As long as the student provides a sound, statistical commentary based on the result they obtained from (a) and (c), awared **2 marks**.