Making Inferences of Two Means

Blake Pappas

11/2/2021

Inference on Two Means

Before beginning an analysis that compares two means, you might want to think through these questions:

1. Are the data paired or independent samples?
2. Would a test or a confidence interval be a more appropriate approach?
3. If the data are two independent samples, is it more reasonable to assume an equal variances or unequal variances model?

When it comes to question 2, in many practical applications, it might be your choice. Testing is great if your primary interest is in whether the data provide strong evidence that the means of the two groups is not equal to a single, specific value. If you simply want to learn about how different the means are, a confidence interval will provide more information.

Also, remember that a hypothesis test can be viewed as an "inversion" of a confidence interval. In the context of inference on two means, this means that if the $(1-\alpha)\cdot 100\%$ CI for $\mu_1 - \mu_2$ includes some value, D_0, then we would fail to reject $H_0 : \mu_1 - \mu_2 = D_0$ against the two-sided alternative. The CI and the test ultimately provide the same information.

In R, if you want to use a D_0 value other than 0 for a two independent sample t test, use the "mu =" option. Here's an example:

```
# Simulate y1 and y2 from Two Normal Distributions
n1 <- 7; n2 <- 9 # Sample sizes in each group
y1 <- rnorm(n1, 2.5, 1)
y2 <- rnorm(n2, 1, 1)

# Test H0: mu1 - mu2 = 1 vs. HA: mu1-mu2 =/= 1
t.test(y1, y2, mu = 1, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  y1 and y2
## t = 1.0441, df = 13.84, p-value = 0.3143
## alternative hypothesis: true difference in means is not equal to 1
## 95 percent confidence interval:
##  0.5664075 2.2544321
## sample estimates:
## mean of x mean of y
## 2.3490570 0.9386373
```

Example: Diabetes

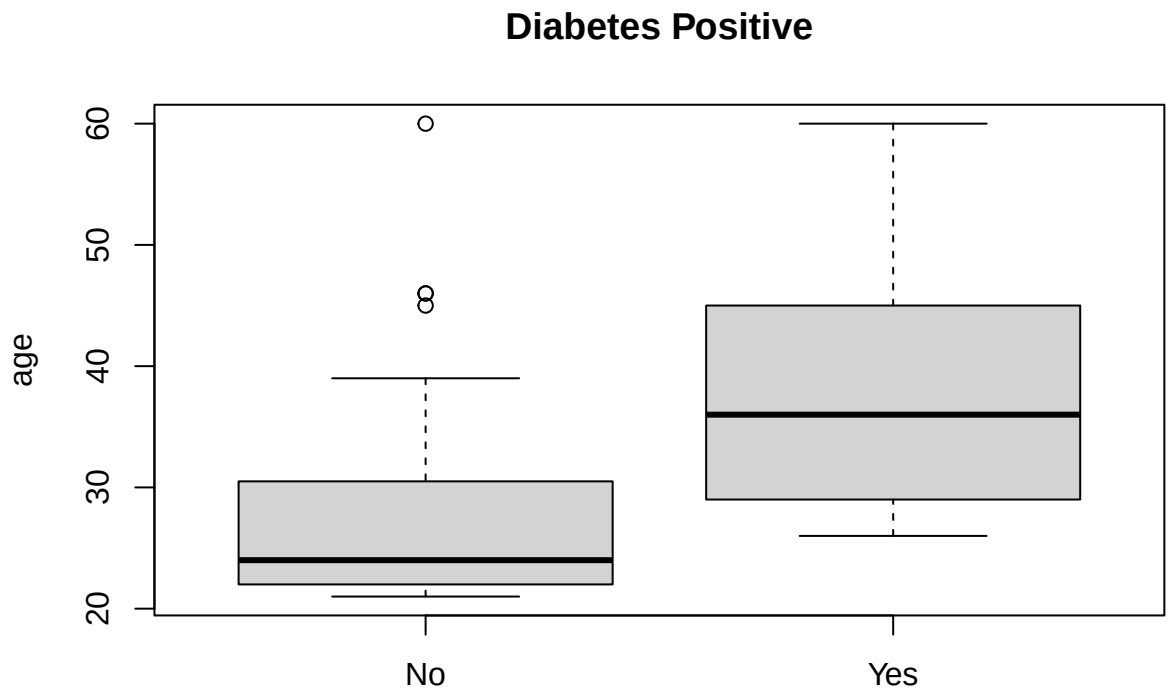
The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on diabetes. A population of several hundred Pima Indian women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including number of pregnancies, glucose level, blood pressure, skin fold thickness, body mass index, diabetes pedigree and age. A random sample (n = 44) of the data from this study are contained in the file `diabetes_sm.csv`.

a. Compare the ages of women who tested negative for diabetes and those who tested positive by making side-by-side boxplots of ages in the two groups. Also compare the mean, median, and standard deviation of "age" across the two groups. Use the variables "age" and "diabetes." Based on these descriptive summaries, does age appear to be associated with result of the diabetes test?

```
diab <- read.csv("diabetes_sm.csv")

diabetes_positive <- subset(diab, diab$diabetes == "Yes")
diabetes_negative <- subset(diab, diab$diabetes == "No")

boxplot(diab$age ~ diab$diabetes, main = "Diabetes Positive", xlab = "", ylab = "age")
```




```
diabetes_positive <- subset(diab, diab$diabetes == "Yes")
diabetes_negative <- subset(diab, diab$diabetes == "No")

mean(diabetes_positive$age)
```

```
## [1] 38.38462
```

```
mean(diabetes_negative$age)
```

```
## [1] 28.45161
```

```
median(diabetes_positive$age)
```

```
## [1] 36
```

```
median(diabetes_negative$age)
```

```
## [1] 24
```

```
sd(diabetes_positive$age)
```

```
## [1] 10.39662
```

```
sd(diabetes_negative$age)
```

```
## [1] 9.804212
```

Answer: Yes, age does appear to be associated with the result of the diabetes test. The higher the age of an individual, the more likely they are to have diabetes.

b. Let μ_1 represent the mean age among women who tested positive for diabetes and let μ_2 represent the mean age among women who tested negative. Find a 95% confidence interval for $\mu_1 - \mu_2$.

Answer: The confidence interval is (2.915219, 16.95079)

c. Using the same definitions of μ_1 and μ_2 as in part (b), test the hypotheses $H_0 : \mu_1 - \mu_2 = 0; H_A : \mu_1 - \mu_2 \neq 0$. Use $\alpha = 0.05$.

(Before you perform the test, use the confidence interval in part (b) to predict in your mind whether H_0 will be rejected.) Report the test statistic, p-value, and conclusion.

```
t.test(diab$age ~ diab$diabetes, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  diab$age by diab$diabetes
## t = -2.9399, df = 21.427, p-value = 0.007715
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -16.950786  -2.915219
## sample estimates:
##  mean in group No mean in group Yes
##          28.45161          38.38462
```

Answer: Looking at the hypothesis test above, the test statistic is -2.9399 and the p-value is 0.007715, which means we reject the null hypothesis. There is sufficient evidence which suggests that the means are different.

Example: Tomato Plants

The file `tomato.csv` contains data on the heights of ten tomato plants (in cm), grown under a variety of soil pH conditions and randomly sampled from their respective plots. Each plant was measured twice. During the first measurement, each plant's height was recorded and a reading of soil pH was taken. During the second measurement only plant height was measured, because it is assumed that pH levels did not vary much from measurement to measurement.

A researcher wants to learn how much the plants grew on average between the first and second measurements.

a. Provide a short descriptive analysis that addresses the researcher's question. (Make a relevant plot and calculate a few relevant summary statistics.)

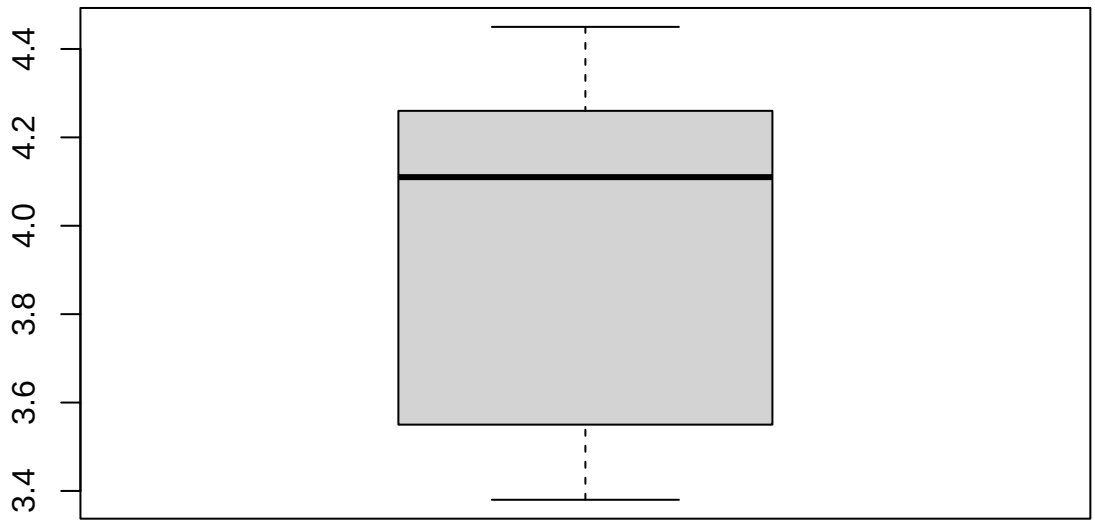
```
tomato <- read.csv("tomato.csv")

tomato$differences <- tomato$height2 - tomato$height1

summary(tomato$differences)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.380   3.635   4.110   3.991   4.253   4.450
```

```
boxplot(tomato$differences)
```



Answer: The tomato's height grows by an average of 3.991 and median of 4.110, with an IQR of 0.618 and overall range of 1.070.

b. Find a 90% confidence interval for the average growth between the first and second measurements.

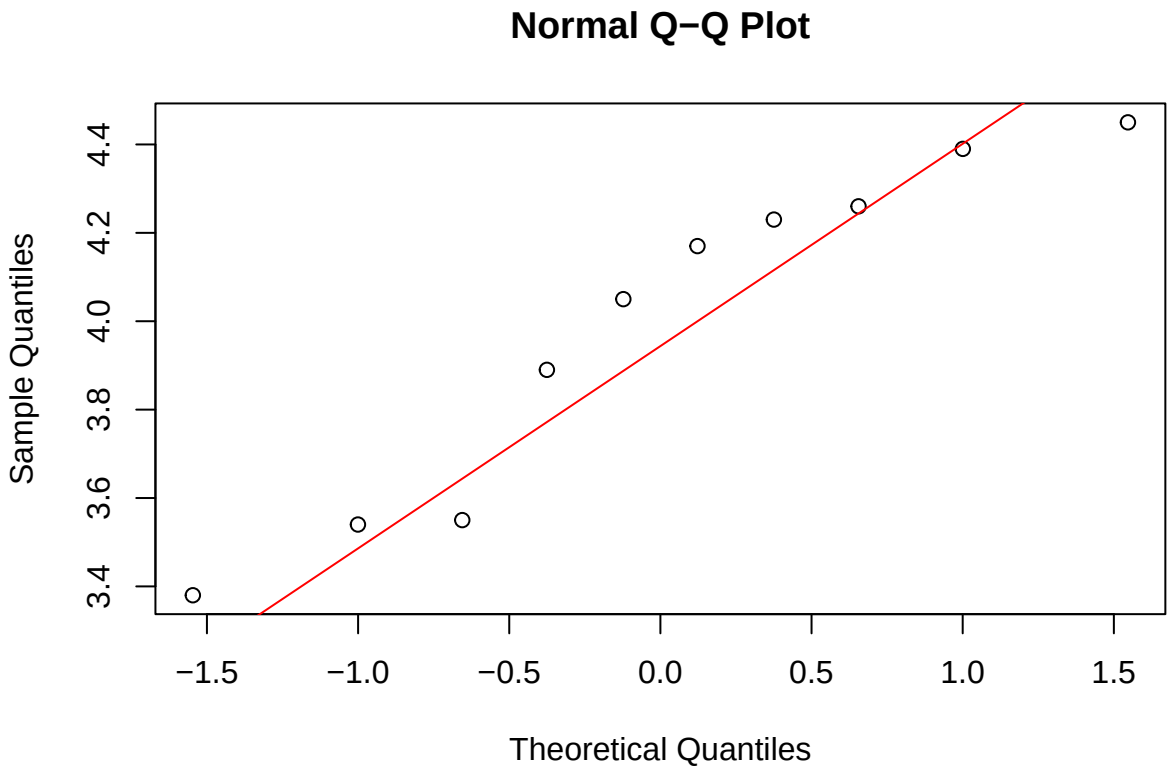
```
t.test(tomato$differences, conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  tomato$differences
## t = 33.001, df = 9, p-value = 1.06e-10
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  3.769311 4.212689
## sample estimates:
## mean of x
##     3.991
```

Answer: The confidence interval for the average growth between the first and second measurements is (3.769311, 4.212689).

c. Check the assumptions of the statistical model using a normal quantile plot. Does the model seem to be a good approximation for the data?

```
qqnorm(tomato$differences)
qqline(tomato$differences, col = 'red')
```



Answer: Please see above for the statistical model representing the average tomato growth. This model looks to be a good approximation of the data, as is evidenced by the curvature in the data points.

Example: Oxygen

The Department of Natural Resources received a complaint from recreational fisherman that a community was releasing sewage into the river where they fish. These types of releases lower the level of dissolved oxygen in the river and hence cause damage to the fish residing in the river. An inspector from the DNR collective fifteen water samples from locations on the river upstream from the community and fifteen samples are selected from locations downstream from the community. The dissolved oxygen readings in parts per million (ppm) were recorded at each location.

The R code below reads the measurements into two objects, "upstream" and "downstream".

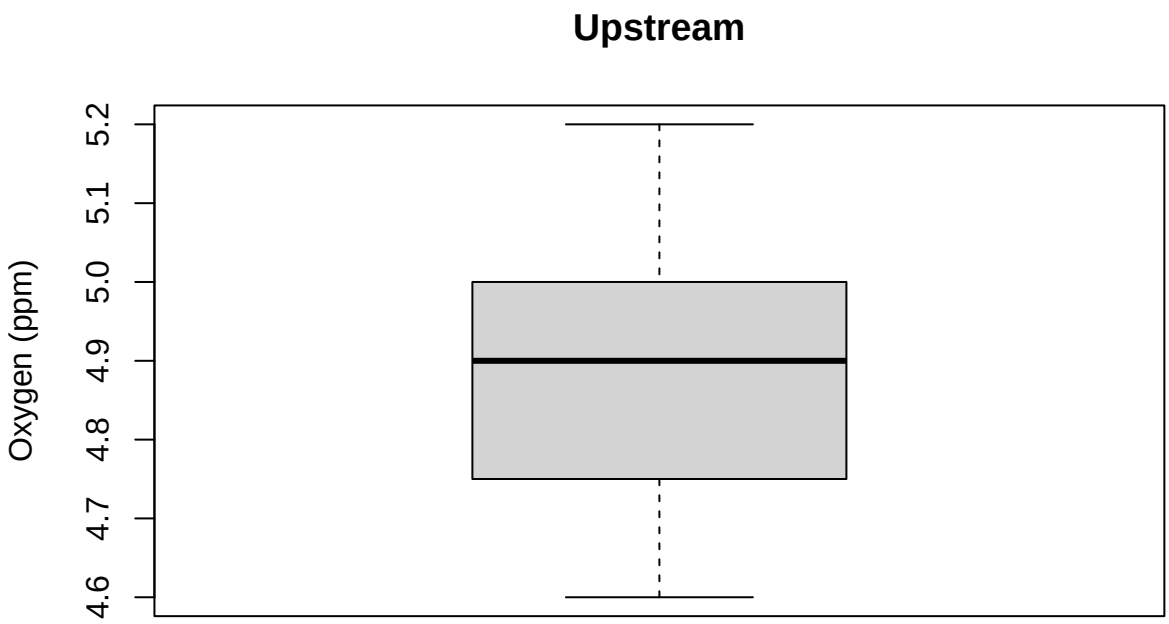
```
upstream <- c(5.2, 4.8, 5.1, 5, 4.9, 4.8, 5, 4.7, 4.7, 5, 4.6, 5.2, 5, 4.9, 4.7)
downstream <- c(3.2, 3.4, 3.7, 3.9, 3.6, 3.8, 3.9, 3.6, 4.1, 3.3, 4.5, 3.7, 3.9, 3.8)
```

a. Are the upstream and downstream measurements paired or independent samples?

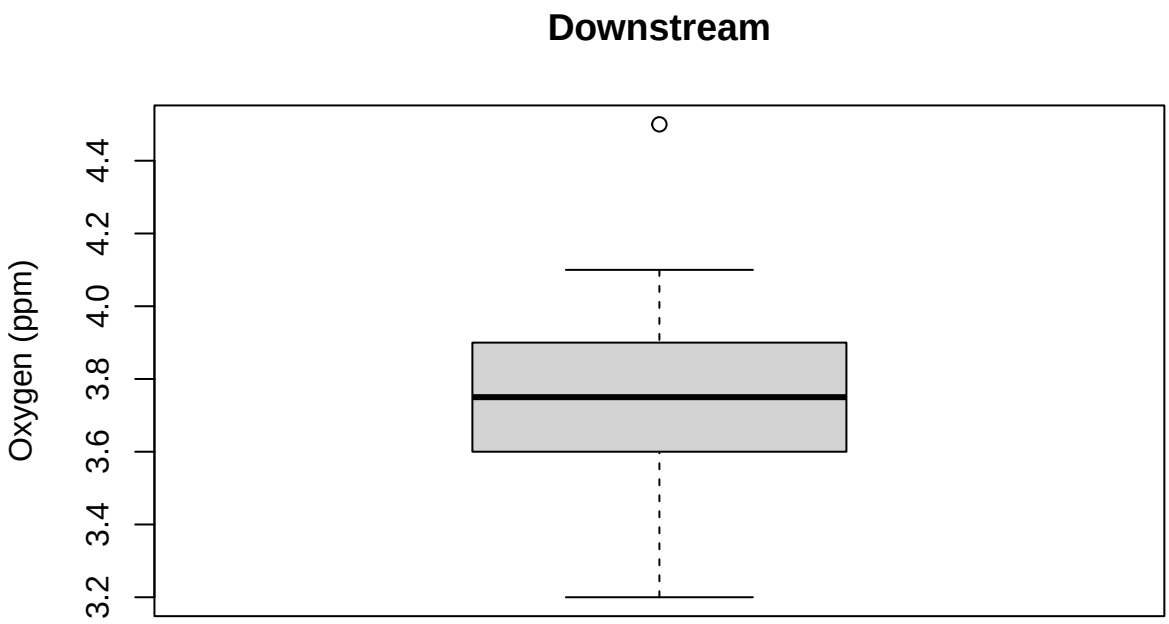
Answer: The upstream and downstream measurements paired samples because they come from the same river.

b. Make a plot to visually investigate whether the dissolved oxygen readings tend to be lower in the downstream or upstream locations.

```
boxplot(upstream, main = "Upstream", xlab = "", ylab = "Oxygen (ppm)")
```




```
boxplot(downstream, main = "Downstream", xlab = "", ylab = "Oxygen (ppm)")
```



Answer: Based on the boxplots above, downstream oxygen levels tend to be lower than that of upstream oxygen levels.

c. The DNR will intervene if the average dissolved oxygen levels upstream is more than 0.5 ppm greater than the average dissolved oxygen levels upstream. Do the data provide strong evidence that the difference in means is greater than 0.5? Report the hypotheses, test statistic, p-value, and conclusion.

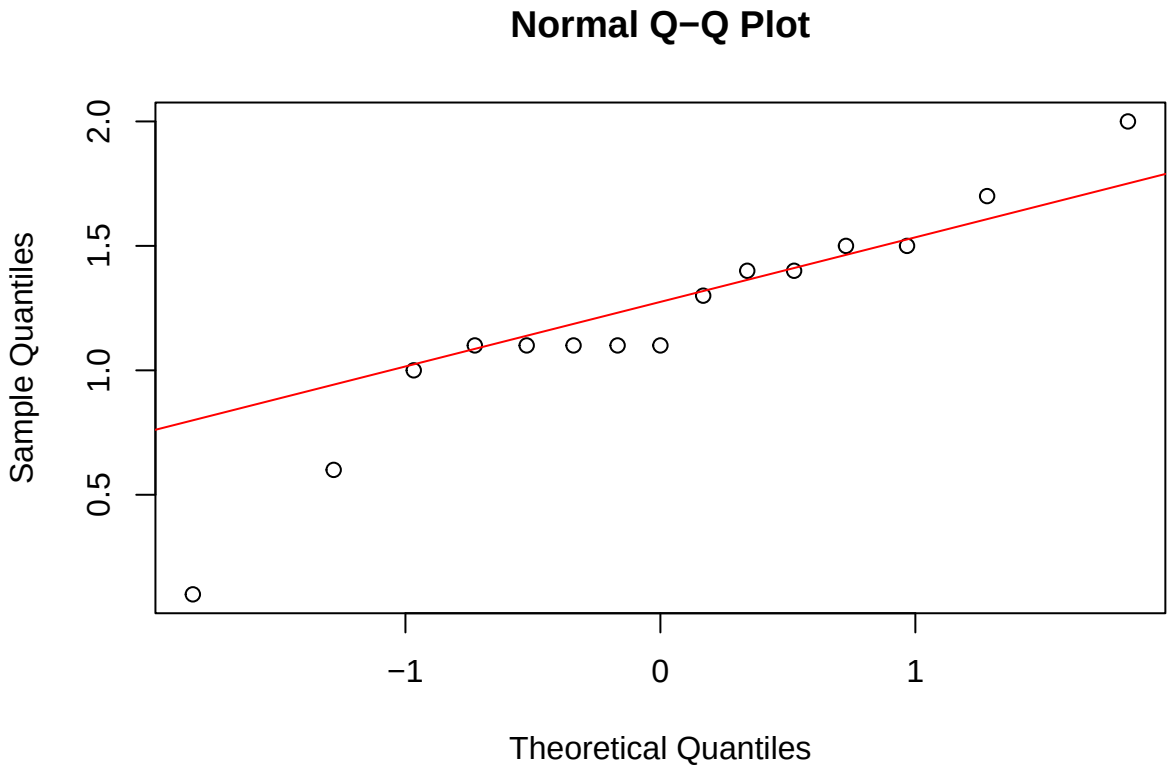
```
t.test(upstream - downstream, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  upstream - downstream
## t = 10.355, df = 14, p-value = 3.025e-08
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.9958964       Inf
## sample estimates:
## mean of x
##       1.2
```

Answer: H0: $\mu_1 - \mu_2 = 0$. HA: $\mu_1 - \mu_2 > 0.5$. The test statistic is 10.355 and the p-value is 3.025e-08. Therefore, the data does not provide strong evidence that the difference in means is greater than 0.5.

d. Check the normality assumptions of the statistical model using quantile plots. Does the model seem to be a good approximation for the data?

```r
qqnorm(upstream - downstream)
qqline(upstream - downstream, col = 'red')
```



Answer: The model does seem to be a good approximation for the data, as is evidenced by the curbature in the data points.

Example: More Tomatoes

Revisit the tomato data in Example 2.

a. Find a 95% confidence interval for the mean pH level for the population of tomatoes.

```r
t.test(tomato$ph, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  tomato$ph
## t = 9.9049, df = 9, p-value = 3.874e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.471481 5.526519
## sample estimates:
```

```
## mean of x
##     4.499
```

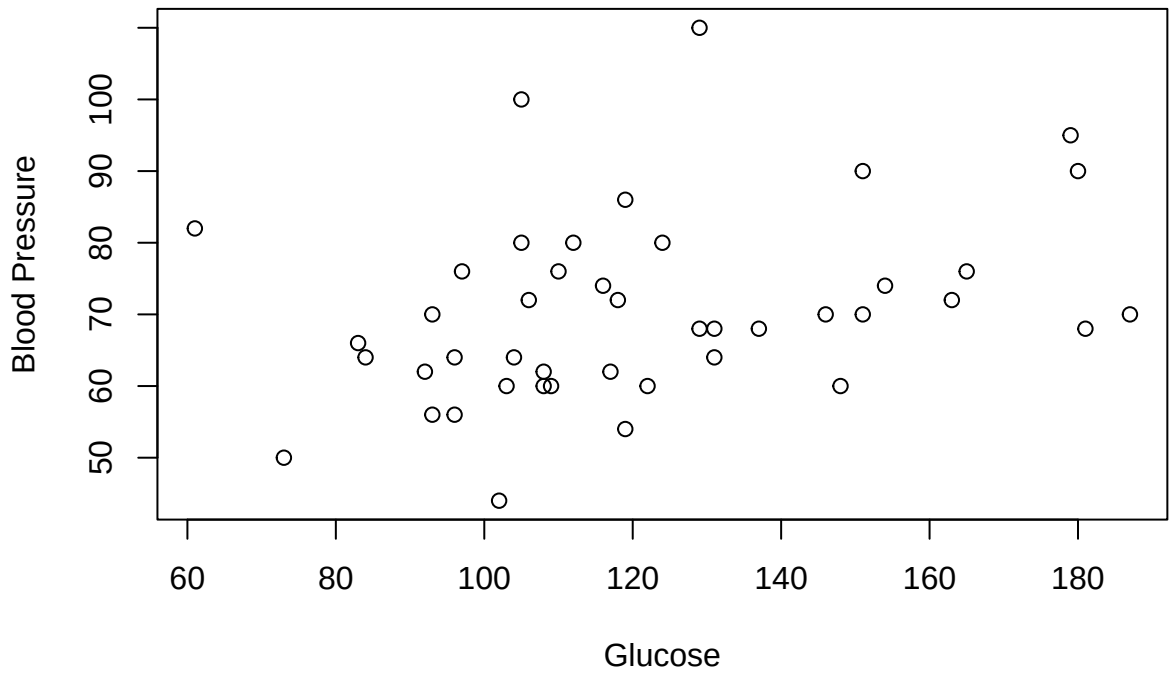
Answer: The confidence interval for the mean pH level for the population of tomatoes is (3.471481, 5.526519).

Example: More Diabetes

Revisit the diabetes data.

a. Provide a descriptive plot and summary statistic that describes the relationship between glucose (glu) and diastolic blood pressure (bp). Does there appear to be a strong association between the two variables?

```
plot(diab$glu, diab$bp, xlab = "Glucose", ylab = "Blood Pressure")
```




```
cor(diab$glu, diab$bp)
```

```
## [1] 0.3216093
```

```
summary(diab$glu)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    61.0   102.8   116.5   121.3   139.2   187.0
```

```
summary(diab$bp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.00   62.00   69.00   70.57   76.00  110.00
```

Answer: Based on the summary statistics, correlation coefficient, and plot, there does not appear to be a strong association between the glucose levels and blood pressure.

b. Create a binary variable that indicates whether an observation has `npreg` (number of pregnancies) greater than 3. Is there an association between this new variable and `diabetes`? Use descriptive methods and/or inferential methods to answer the question.

Answer: Based on the plot above, there does not appear to be an association between this new variable and diabetes.