t-Based Confidence Intervals

Math 445, Spring 2017

One Sample t-based Intervals

Example

The data set `MnGroundwater` contains measurements on the water quality of 895 randomly selected wells in Minnesota. We need to construct a 95% confidence interval for the mean alkalinity level.

```
library(resampledata)
```

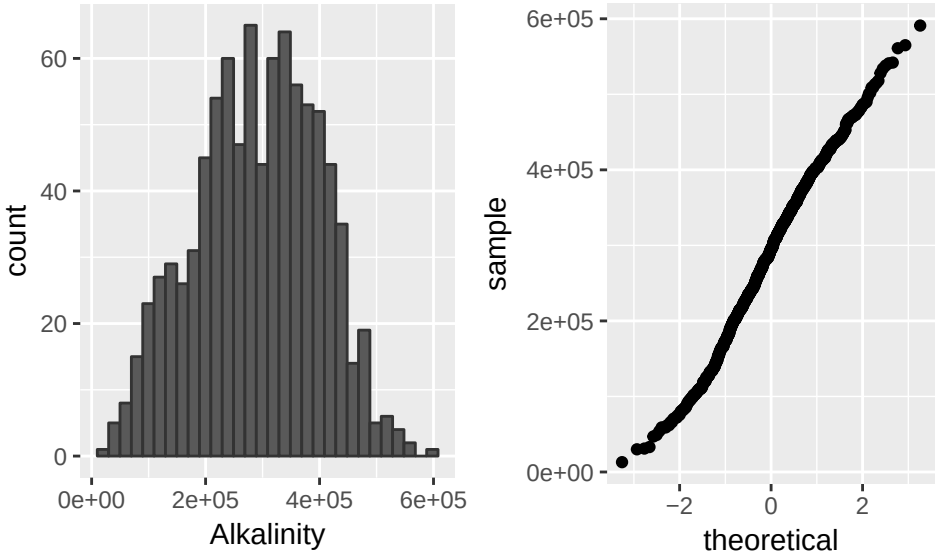
```
##
## Attaching package: 'resampledata'
```

```
## The following object is masked from 'package:datasets':
##
##     Titanic
```

Let's begin by plotting the data:

```
ggplot(data = MnGroundwater) +
  geom_histogram(mapping = aes(x = Alkalinity), colour = "gray20")

ggplot(data = MnGroundwater) +
  stat_qq(mapping = aes(sample = Alkalinity))
```



What do these plot reveal about the assumptions required for the t-based confidence interval?

What about the other necessary assumptions?

Two-sided confidence intervals

Assuming that all of the assumption are upheld, we can build a t-based confidence in R using the `t.test` function. To see the help file, run the command `?t.test`.

```r
t.test(MnGroundwater$Alkalinity, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  MnGroundwater$Alkalinity
## t = 80.272, df = 894, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  283575.6 297789.8
## sample estimates:
## mean of x
##  290682.7
```

Notice that this function does more than you asked for—it conducts a hypothesis test as well! Since `t.test` returns a list, you can easily extract only the element of interest

```r
t.test(MnGroundwater$Alkalinity, conf.level = 0.95)$conf.int
```

```
## [1] 283575.6 297789.8
## attr(,"conf.level")
## [1] 0.95
```

How do we interpret this interval?

One-sided confidence intervals

There may be situations where you are only interested in an upper or lower bound. In these situations we can easily adapt the two-sided t interval in R using the `alternative` argument:

```r
t.test(MnGroundwater$Alkalinity, conf.level = 0.95, alternative = "less")$conf.int
```

```
## [1]     -Inf 296645.2
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(MnGroundwater$Alkalinity, conf.level = 0.95, alternative = "greater")$conf.int
```

```
## [1] 284720.1      Inf
## attr(,"conf.level")
## [1] 0.95
```

Two Sample t-based Intervals

Example

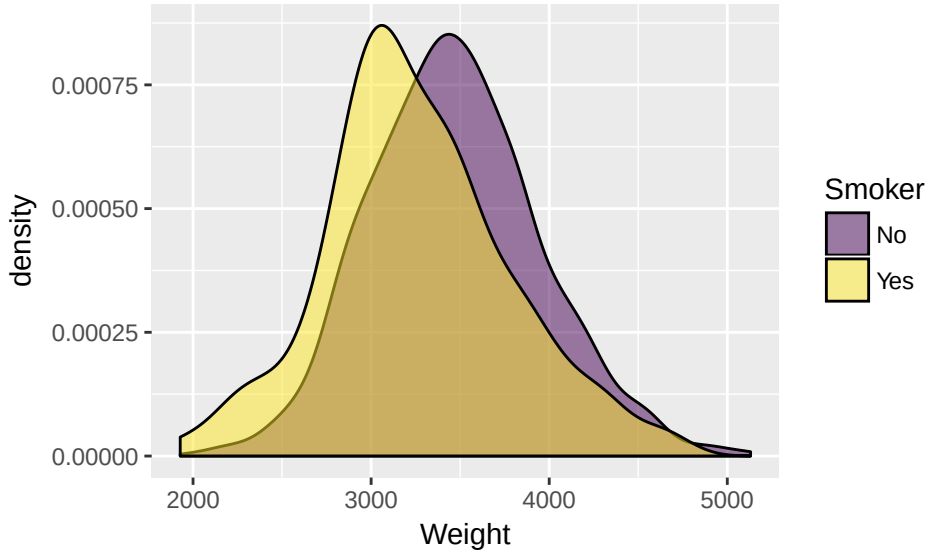
The birth weight of a baby is of interest to health officials since many studies have shown possible links between this weight and conditions in later life, such as obesity or diabetes. Researchers look for possible relationships between the weight of a baby and the age of the mother, or whether or not she smoked cigarettes or drank alcohol during her pregnancy. We will investigate data consisting of a random sample of 1009 babies born in North Carolina during 2004.

Let's begin by plotting the data:

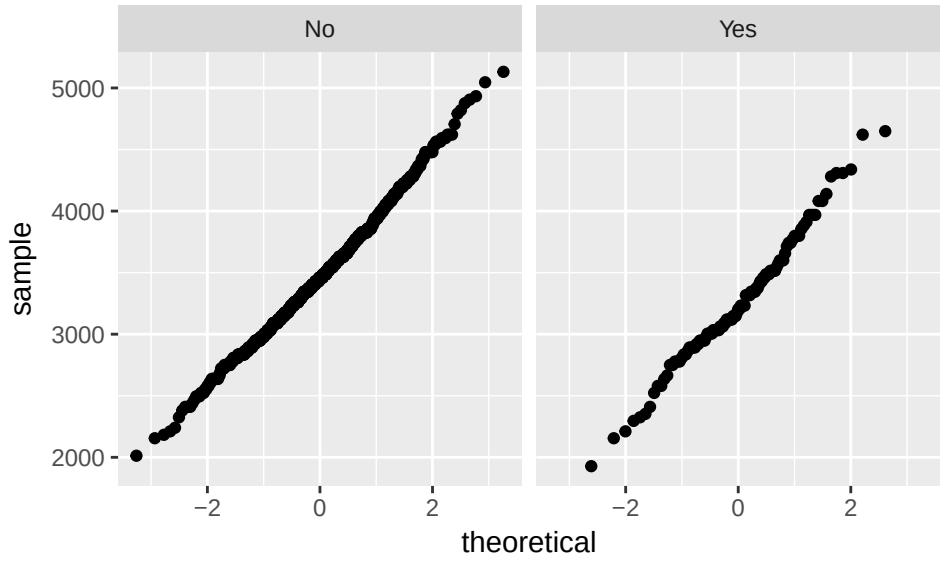
```r
library(viridis)
```

```
## Loading required package: viridisLite
```

```
ggplot(data = NCBirths2004) +
  geom_density(mapping = aes(x = Weight, fill = Smoker), alpha = 0.5) +
  scale_fill_viridis(discrete = TRUE)
```




```
ggplot(data = NCBirths2004) +
  stat_qq(mapping = aes(sample = Weight)) +
  facet_wrap(~Smoker, ncol = 2)
```



What do these plot reveal about the assumptions necessary for t-based inference?

Welch's t procedure

Assuming that all of the assumption are upheld, we can again build a t-based confidence in R using the `t.test` function. First, let's assume that the variances are not equal; thus, we need to use Welch's t procedure and the Welch-Satterthwaite degrees of freedom.

`t.test` requires a vectors for each group when conducting the two-sample t procedures, so we must first create these vectors via subsetting:

```
s_weight  <- subset(NCBirths2004, select = Weight, subset = Smoker == "Yes", drop = TRUE)
ns_weight <- subset(NCBirths2004, select = Weight, subset = Smoker == "No", drop = TRUE)
```

Now we can use the `t.test` function.

```
t.test(ns_weight, s_weight, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  ns_weight and s_weight
## t = 4.1411, df = 134.01, p-value = 6.08e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  112.3161 317.6881
## sample estimates:
## mean of x mean of y
##  3471.912  3256.910
```

How do we interpret this interval?

Pooled t procedure

If there is reasonable evidence that the variances between the two groups are equal, then we should use the pooled sample variance with the t procedure. To do this, add the argument `var.equal = TRUE` to `t.test`.

```
t.test(ns_weight, s_weight, conf.level = 0.95, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  ns_weight and s_weight
## t = 4.4215, df = 1007, p-value = 1.087e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  119.5817 310.4226
## sample estimates:
## mean of x mean of y
##  3471.912  3256.910
```

How does this interval differ from the interval obtained using Welch's t procedure.