Bob the builder (One sample t-test)

We wish to test the hypotheses of getting the job done. Does Bob the builder get the job done 2 times per day or not?

$H_0 : \mu = 2$
$H_1 : \mu \neq 2$

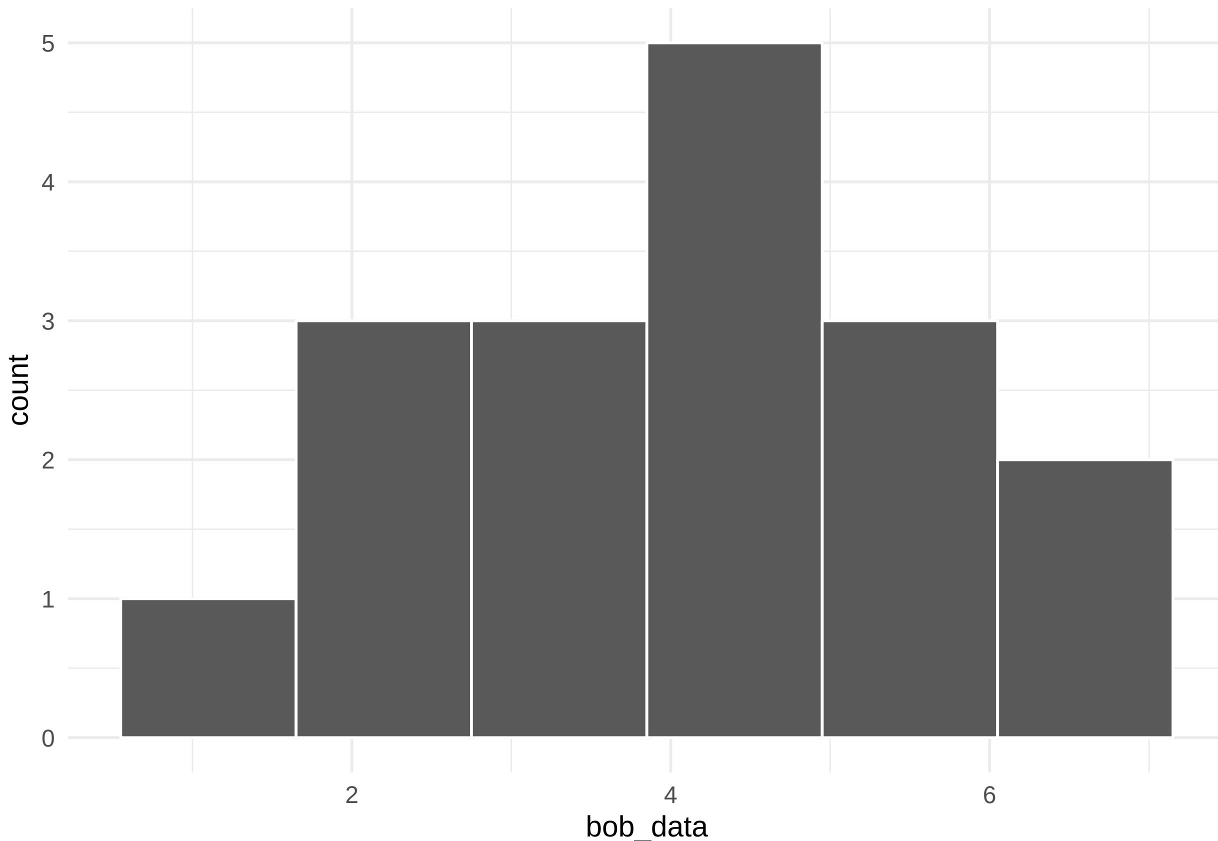
```
mu_0 <- 2
```

We have a random sample of independent observations.

```
bob_data <- c(4.75, 4.4, 3.8, 5.2, 4.2, 4.7, 5.12, 4.9, 6, 2, 2.3, 1.5, 2.2, 3.8, 3.7, 6.5, 6.2)
```

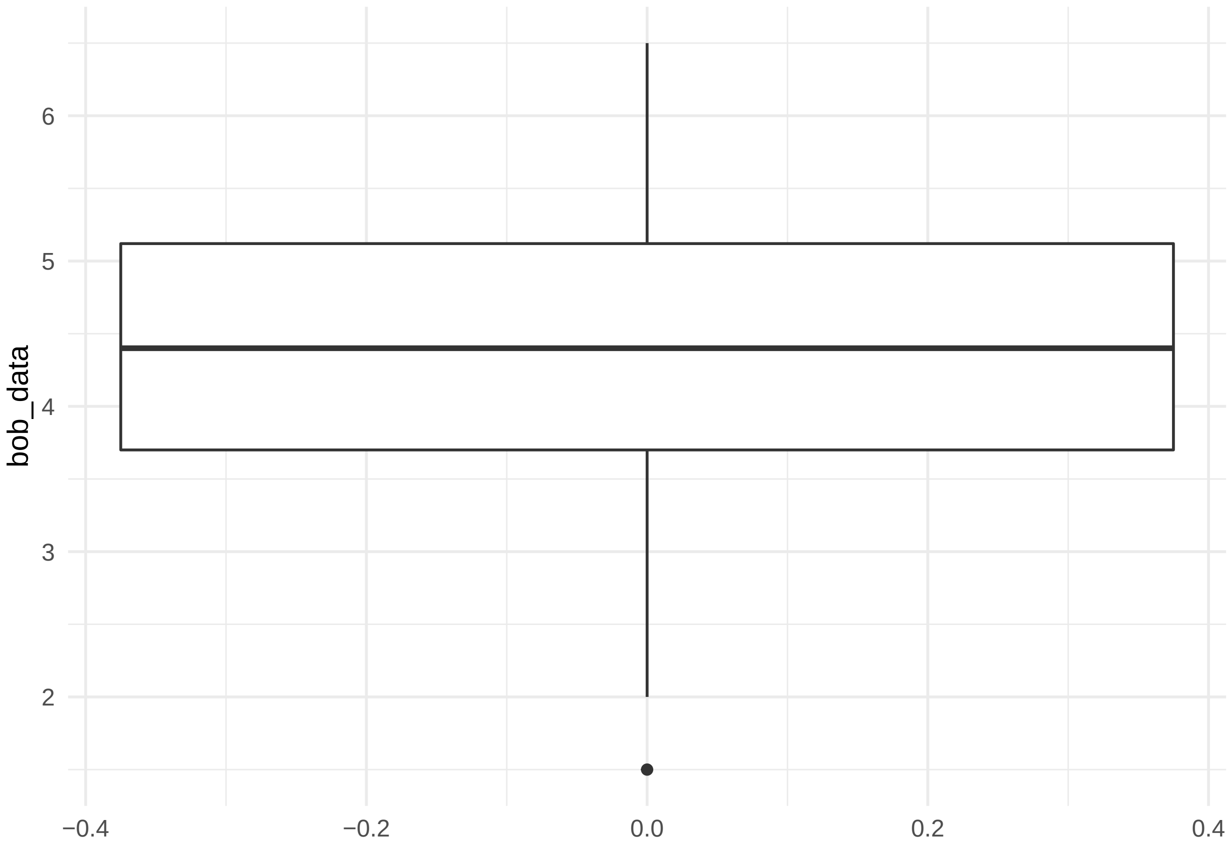
The histogram doesn't look too bad? We have enough data (n=17).

```
library(ggplot2)
ggplot(data.frame(bob_data=bob_data), aes(x=bob_data)) +
  geom_histogram(binwidth=1.1, col="white", lwd=0.5) +
  theme_minimal()
```



There is one outlier.

```
ggplot(data.frame(bob_data=bob_data), aes(y=bob_data)) +
  geom_boxplot() +
  theme_minimal()
```



A t-test is robust, so with caution from above, we'll proceed.

"By Hand" Calculation

Meaning: Use `R` like it is a simple calculator.

```
# * THIS IS BY HAND
n         <- length(bob_data)
x_bar     <- mean(bob_data)
sample_sd <- sd(bob_data)
c(n=n, x_bar=x_bar, sample_sd=sample_sd)
```

```
##         n     x_bar sample_sd
## 17.000000  4.192353  1.495157
```

We are using $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$. Check out what t equals.

```
# * THIS IS CONTINUING THE BY HAND CALCULATION
t     <- (x_bar-mu_0) / (sample_sd/sqrt(n))
t
```

```
## [1] 6.045722
```

By definition, we have degrees of freedom as 1 minus the number of observations.

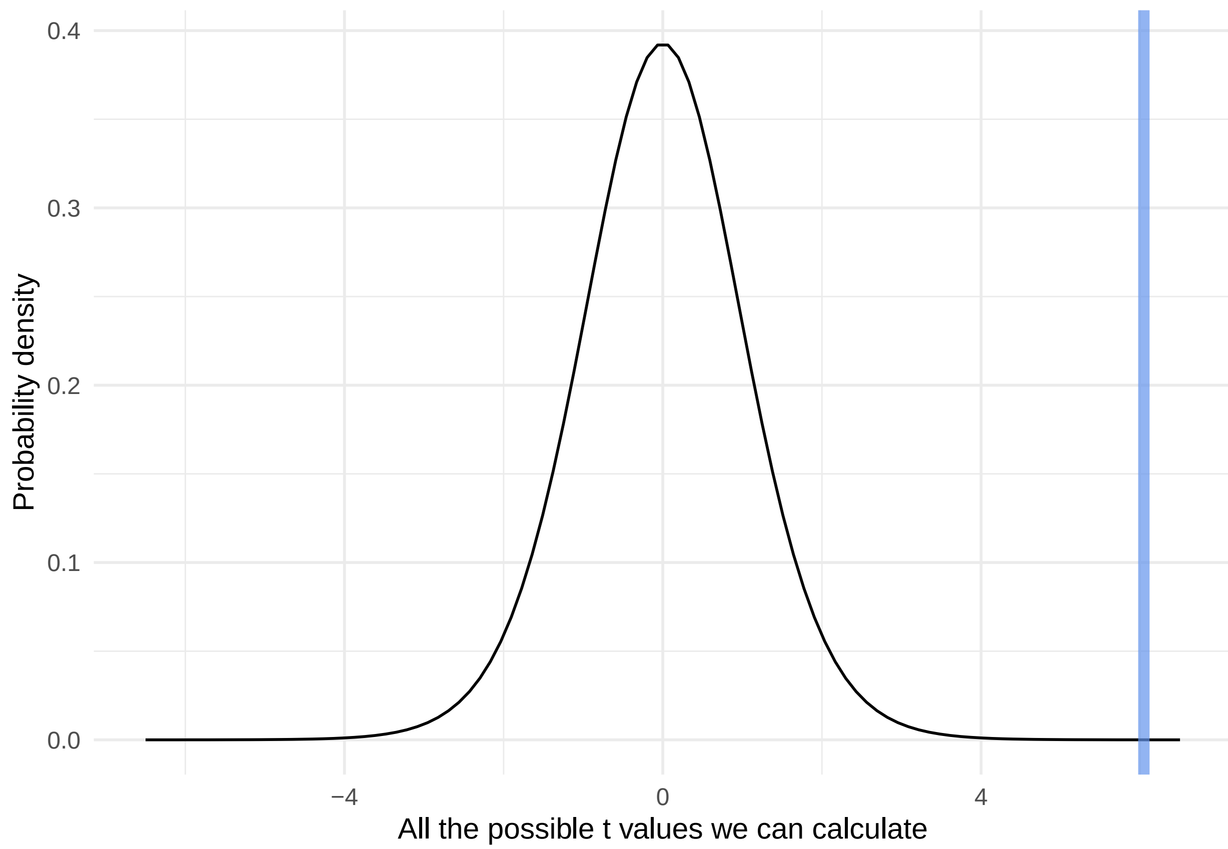
```
# * THIS IS CONTINUING THE BY HAND CALCULATION
df    <- n-1
df
```

```
## [1] 16
```

Let's see where t lands on our distribution. I am plotting a t-distribution with $df = n - 1 = 16$.

```
# * THIS IS THE T-DISTRIBUTION WE ARE COMPARING AGAINST
x  <- seq(-6.5, 6.5, length=100)
hx <- dt(x, df=n-1)
t_dist <- data.frame(cbind(x,hx))

ggplot(t_dist, aes(x=x, y=hx)) +
  geom_line() +
  geom_vline(xintercept=t, col="cornflowerblue", lwd=2, alpha=0.7) +
  xlab("All the possible t values we can calculate") +
  ylab("Probability density") +
  theme_minimal()
```



Can you guess what our p-value will be? (Big? Small?) We're going to take the area of being above the blue line on the above distribution as our p-value. (The probability of rejecting H_0 given thatn H_0 is actually the truth.)

```
# * THIS IS CONTINUING THE BY HAND CALCULATION
p_val <- 2*(1 - pt(q=t, df=df))
p_val
```

```
## [1] 1.699117e-05
```

Look at slides to see interpretation of p-value!

Also, question: Would the corresponding confidence interval include or not include $\mu_0 = 2$?

"Using R" Calcuation

Meaning: Use more than just simple `R` functions.

```r
# * THIS IS "USING R"
test <- t.test(x=bob_data, alternative="two.sided", mu=2)
test$p.value
```

```
## [1] 1.699117e-05
```