

Lecture 12: One Sample Hypothesis Tests

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020



WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

One Sample Hypothesis Test: Example I



The manufacturer knows that as long as the thickness is greater than 1.5mm, all is good. They are willing to accept some loss, so they really just want to make sure that, on average, the thickness is greater than 1.5mm.

The manufacturer decides to hire you as a consultant. They want your help to answer the question: "is the mean thickness greater than 1.5mm?"

As a statistical wizard, you know to test such a hypothesis, the first thing to do is to setup a null and alternative hypothesis. Recall: the null is always the simplest/least interesting hypothesis. So, if we let μ denote the mean paint thickness, we would like to test the hypothesis

$$H_0 : \mu = 1.5 \quad \text{against} \quad H_A : \mu < 1.5.$$

Next, we have to decide on a reasonable significance level, α . Remember, α is the probability of making a Type I error. I.e. $\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$.

We pick $\alpha = 0.05$. I.e. we decide that it is acceptable to us to rejecting the null, even if it is true, 5% of the time.

One Sample Hypothesis Test: Example I



We collect a sample of 16 engines, and measure the paint thickness. Our first step is to visualize the data:

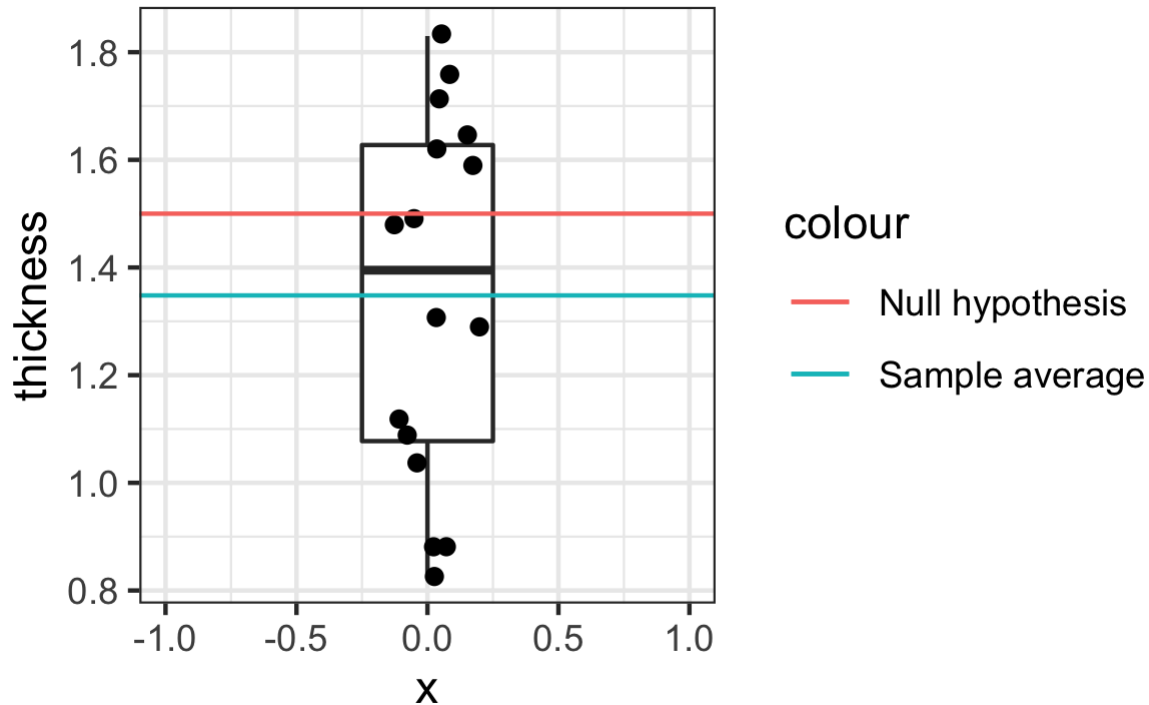
```
library(tidyverse); library(distributions3); theme_set(theme_bw())
paint_thickness <- tibble(
  thickness = c(1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83,
                1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49))

ggplot(data = paint_thickness,
       aes(y = thickness)) +
  geom_boxplot(width = 0.5) +
  geom_jitter(aes(x = 0),
              width = 0.2) +
  geom_hline(aes(yintercept = 1.5,
                 color = "Null hypothesis")) +
  geom_hline(aes(yintercept = mean(thickness),
                 color = "Sample average")) +
  xlim(c(-1,1))
```

One Sample Hypothesis Test: Example I



We collect a sample of 16 engines, and measure the paint thickness. Our first step is to visualize the data:



The question then is: are the two lines so far apart that we do no longer think that $\mu = 1.5$ is reasonable? Does the data seem to suggest we should change our beliefs?

One Sample Hypothesis Test: Example I



The next step is to find a good measure of what we are looking for: difference between the observed (\bar{x}) and the null hypothesis ($\mu = 1.5$) relative to the variation in the data (σ) and sample size (n).

Maybe $T = \frac{\bar{X} - 1.5}{\sigma/\sqrt{16}}$ will work?

We do not know σ , so would need to consider $T = \frac{\bar{X} - 1.5}{s/\sqrt{16}}$. To find out if this

would work, we want to ask: can we find

$P(T \text{ more extreme than } T_{\text{obs}} | H_0 \text{ is true})$? To find this, we need to find the distribution of T **ASSUMING** the null hypothesis is true. I.e. we will pretend $\mu = 1.5$ is the true mean paint thickness. If this is the case,

$$T = \frac{\bar{X} - 1.5}{s/\sqrt{16}} = \frac{\bar{X} - E(\bar{X})}{\widehat{\text{SD}}(\bar{X})}.$$

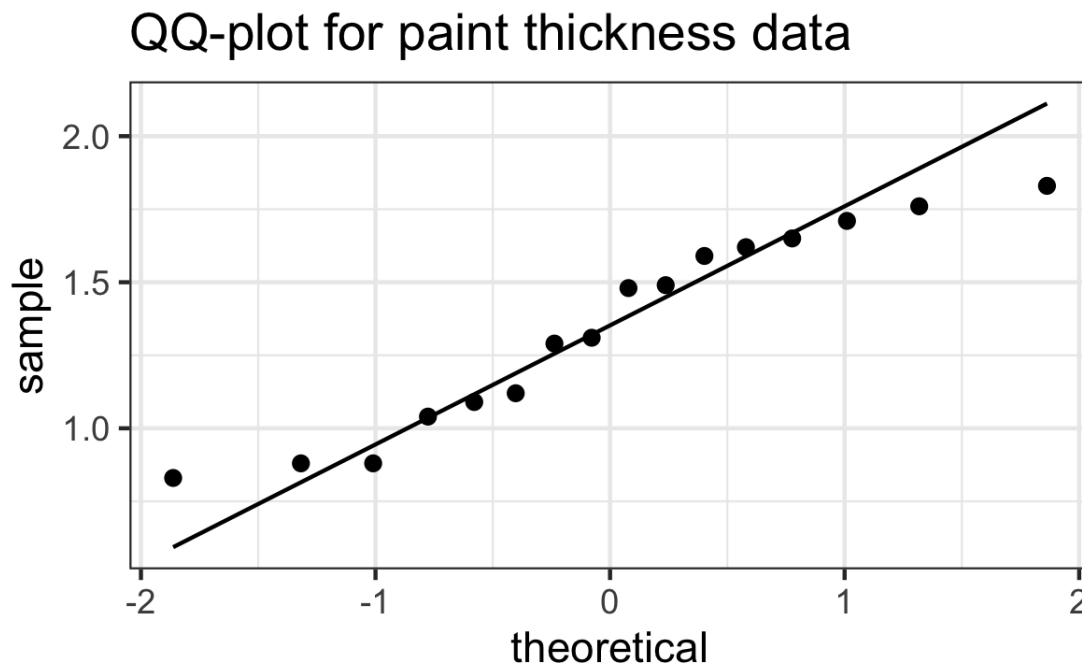
We know that this follows a t -distribution with $n - 1$ degrees of freedom, **IF** \bar{X} is normally distributed.

One Sample Hypothesis Test: Example I



Since $n < 30$, need the data to be normal for $\bar{X} \sim N$. QQ-plot:

```
ggplot(data = paint_thickness, aes(sample = thickness)) +  
  geom_qq() +  
  geom_qq_line() +  
  labs(title = "QQ-plot for paint thickness data")
```



Seems like it is not crazy to think the data is normally distributed.

One Sample Hypothesis Test: Example I



So, we know that $T = \frac{\bar{X} - 1.5}{s/\sqrt{16}} \sim t_{n-1}$ **IF** the null hypothesis $\mu = 1.5$ is true.

Next, calculate the observed value of T :

```
paint_thickness %>%  
  summarize(xbar = mean(thickness),  
            s = sd(thickness),  
            T_obs = (xbar - 1.5)/(s/sqrt(16)))
```

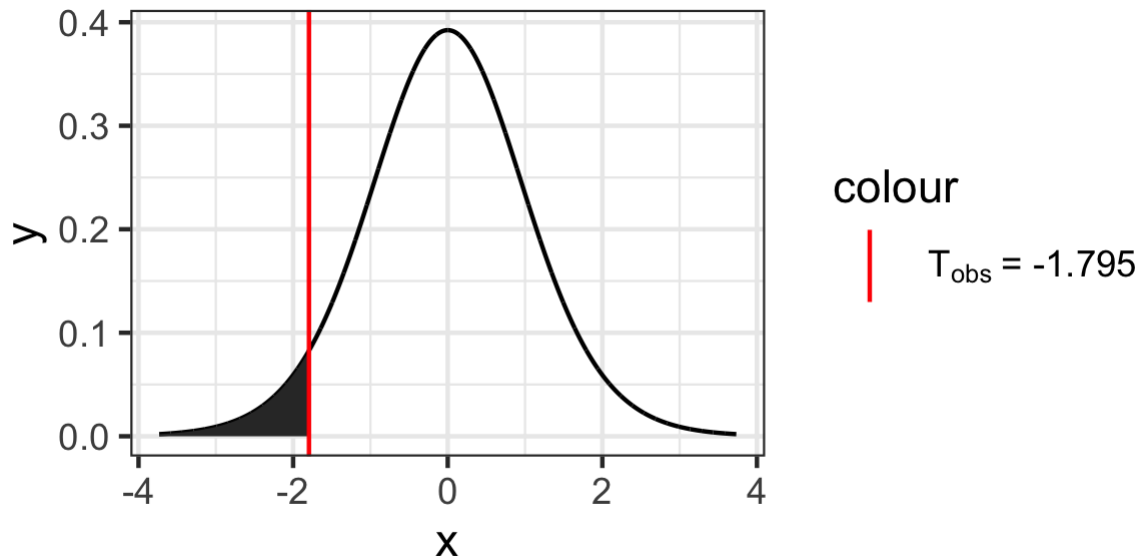
```
## # A tibble: 1 x 3  
##   xbar      s T_obs  
##   <dbl> <dbl> <dbl>  
## 1  1.35 0.339 -1.79
```

Recall, $T_{\text{obs}} \ll 0$ if $\bar{x} \ll 1.5$, and close to 0 if \bar{x} is close to 1.5. So, want to ask: is T_{obs} so small (i.e. very negative) that we do not think the true mean is 1.5, but instead think that $\mu < 1.5$?

One Sample Hypothesis Test: Example I



IF the null is true, $T_{\text{obs}} \sim t_{15}$. So, **IF** the null is true, and T_{obs} is very small, then $P(T_{15} < T_{\text{obs}})$ must be very small. Remember, here "small" means "smaller than α ". We picked 0.05.



One Sample Hypothesis Test: Example I



We find $P(T_{15} < T_{obs}) = P(T_{15} < -1.795)$:

```
cdf(T_15, -1.795)
```

```
## [1] 0.04641306
```

Since this is greater than α , we conclude that $T_{obs} = -1.795$ is NOT crazy small.

I.e. \bar{x} is NOT much smaller than 1.5.

So, we do NOT reject the idea the the true mean is in fact 1.5.

One Sample Hypothesis Test: Example I



Summary:

- Want to test $H_0 : \mu = 1.5$ against $H_A : \mu < 1.5$ using $\alpha = 0.05$.
- Find that we can assume the data are normal from QQ-plot, so \bar{X} is normal.
- **IF** $\mu = 1.5$ is true, then $T = \frac{\bar{X} - 1.5}{s/\sqrt{16}} = \frac{\bar{X} - E(\bar{X})}{\widehat{SD}(\bar{X})} \sim t_{15}$.
- **IF** $\mu = 1.5$, then the probability of something "more extreme" is $P(T_{15} < T_{\text{obs}}) \approx 0.33$.
- **IF** $\mu = 1.5$, then something "more extreme" isn't super unlikely, so our observation is not that unlikely
- Since our observation isn't unlikely **IF** $\mu = 1.5$, we do not reject $H_0 : \mu = 1.5$

One Sample Hypothesis Test: Example I



On second thought, the manufacturer is actually also interested in the mean paint thickness not being TOO large. Thicker layer of paint, more materials used, increased cost.

So, they would like to test the hypothesis

$$H_0 : \mu = 1.5 \quad \text{against} \quad H_A : \mu \neq 1.5.$$

Once again, we pick $\alpha = 0.05$.

Most of the procedure is the same:

- we still ask, if \bar{X} is close enough to 1.5 that we do still believe in the null hypothesis.
- we still answer in terms of the difference compared to the variation:

$$T = \frac{\bar{X} - 1.5}{\sigma/\sqrt{16}}$$

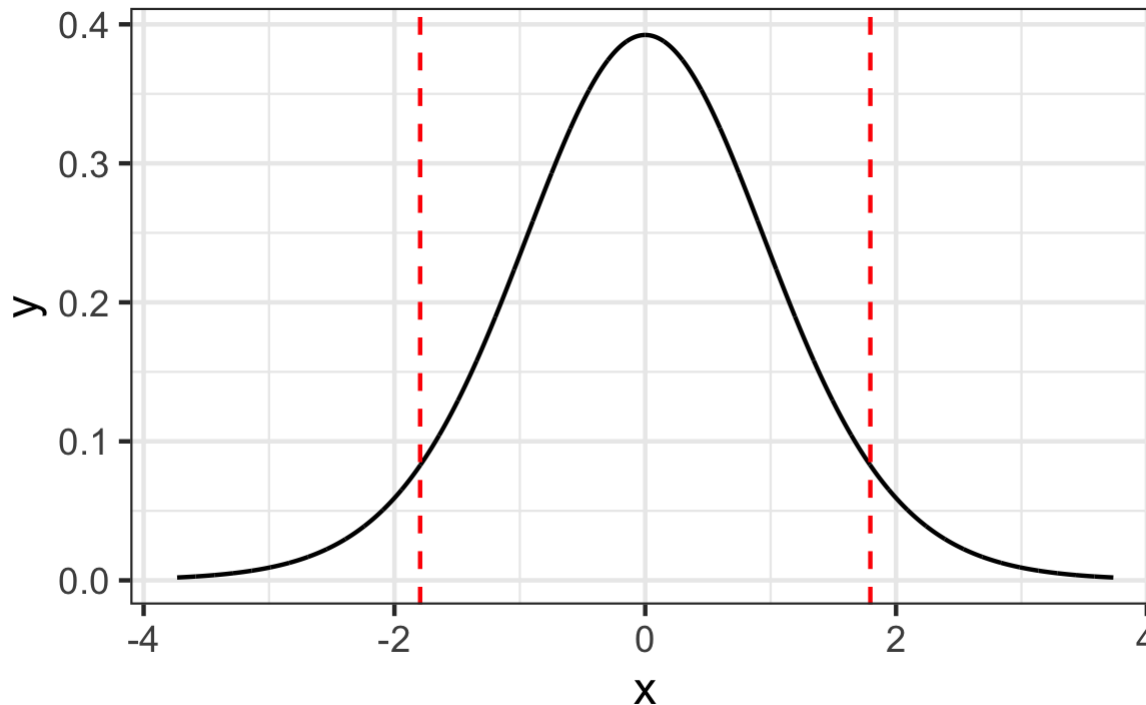
- \bar{X} is still normally distributed (the QQ-plot is the same, so data still normal...), so $T \sim t_{15}$

The difference is in what it means to be "more extreme".

One Sample Hypothesis Test: Example I



First, remember that $T_{\text{obs}} = -1.795$. So, what does it mean to be "more extreme"? Since the alternative is $H_A : \mu \neq 1.5$, "more extreme" goes in both directions: "more extreme" is "further from 0", i.e. outside the dotted lines below:



One Sample Hypothesis Test: Example I



So, the p-value is

$$\begin{aligned} P(\text{more extreme} \mid H_0 \text{ true}) &= P(T < -1.795 \text{ OR } T > 1.795) \\ &= P(T < -1.795) + P(T > 1.795) \\ &= 2 \cdot P(T < -1.795) \end{aligned}$$

Found in R:

```
p_value <- 2*cdf(T_15, -1.795)
p_value
```

```
## [1] 0.09282611
```

Since $p > 0.05$, we do NOT reject the null hypothesis:

- **IF** the null is true, the probability of seeing something "further away" from the null is large
- **IF** the null is true, our observed data are NOT unlikely
- since our observed is NOT unlikely **IF** the null is true, our observed data provides little evidence against the null
- we do NOT reject the null.

One Sample Hypothesis Test: Example II



Recall: Framingham Heart Study. Interested in the true population mean total cholesterol. In particular, is the population mean total cholesterol 230 mg/dL?

To perform a hypothesis test, we setup our null- and alternative hypotheses: we will test

$$H_0 : \mu = 230 \quad \text{against} \quad H_A : \mu \neq 230.$$

This time, we pick $\alpha = 0.01$, i.e. we will only accept making a Type I error 1% of the time.

One Sample Hypothesis Test: Example II



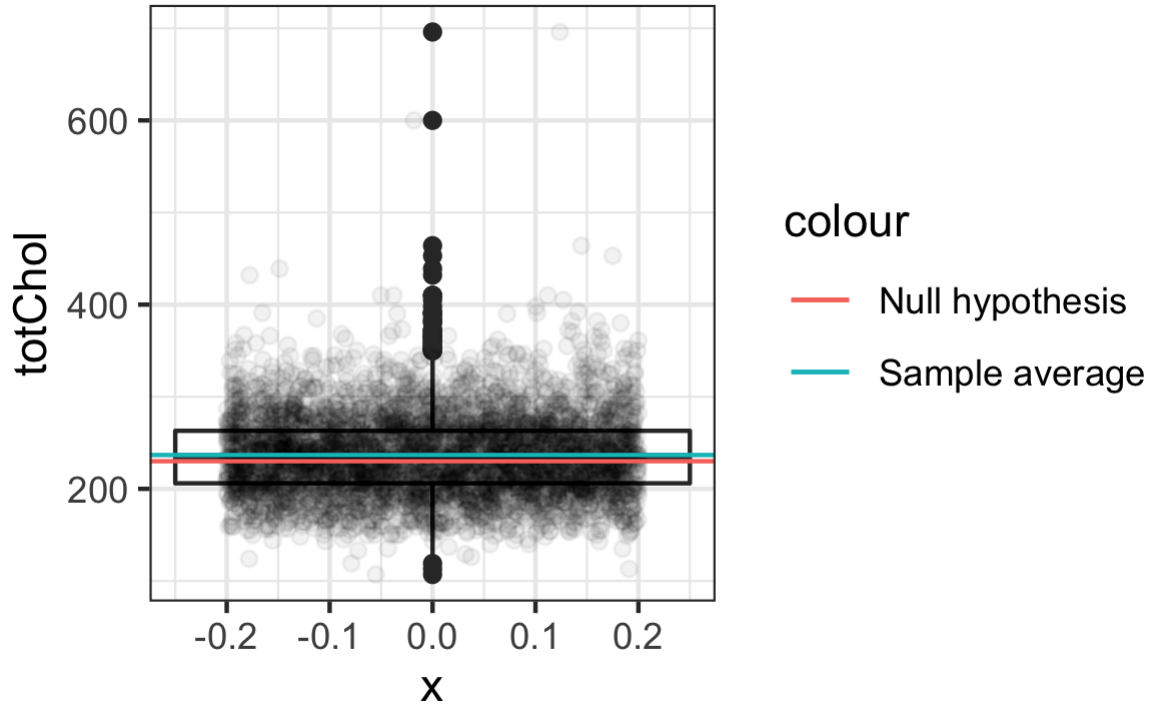
We'll take a look at the data:

```
fram <- read_csv(here::here("csv_data/framingham.csv")) %>%  
  filter(!is.na(totChol))  
  
ggplot(fram,  
       aes(y = totChol)) +  
  geom_boxplot(width = 0.5) +  
  geom_jitter(aes(x = 0),  
             height = 0, width = 0.2, alpha = 0.05) +  
  geom_hline(aes(yintercept = 230,  
                color = "Null hypothesis")) +  
  geom_hline(aes(yintercept = mean(totChol),  
                color = "Sample average"))
```

One Sample Hypothesis Test: Example II



We'll take a look at the data:



One Sample Hypothesis Test: Example II



Once again, we will assess this using $T = \frac{\bar{X} - 230}{\widehat{SD}(\bar{X})} = \frac{\bar{X} - 230}{s/\sqrt{n}}$.

IF the null hypothesis is true, then $E(\bar{X}) = 230$.

Therefore, **IF** the null hypothesis is true,

$$T = \frac{\bar{X} - 230}{s/\sqrt{n}} = \frac{\bar{X} - E(\bar{X})}{\widehat{SD}(\bar{X})} \sim t_{n-1}.$$

Let's find T_{obs} :

```
fram %>%  
  summarize(xbar = mean(totChol),  
            SD = sd(totChol),  
            n = n(),  
            T_obs = (xbar - 230)/(SD/sqrt(n)))
```

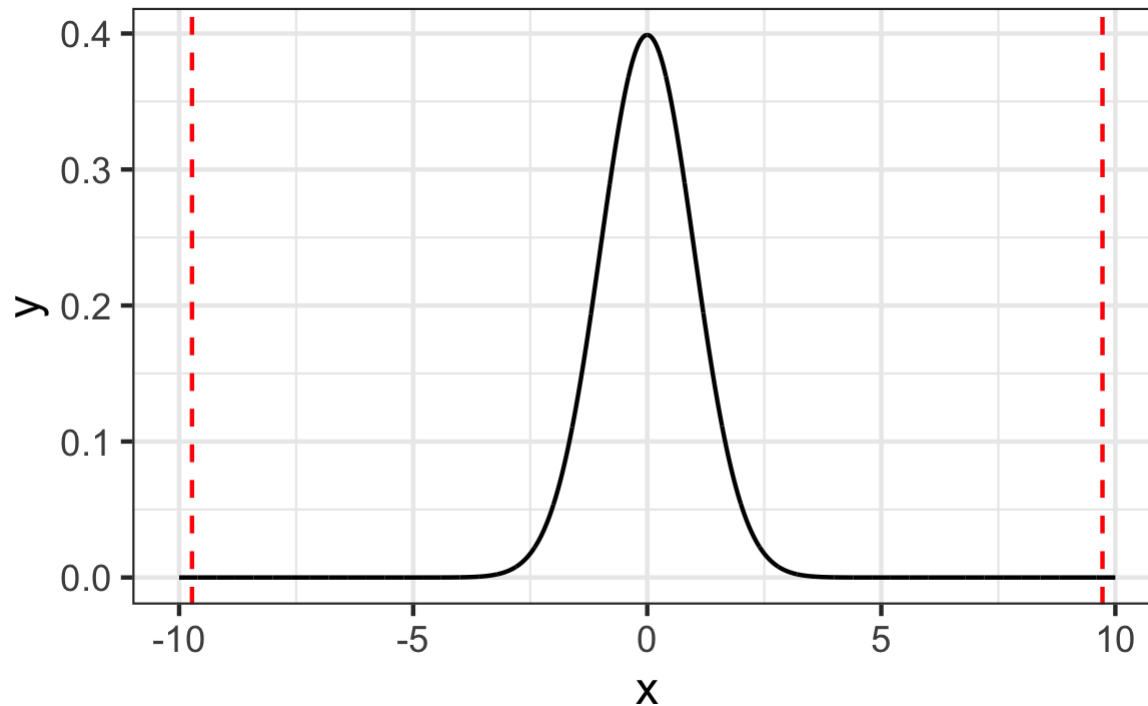
```
## # A tibble: 1 x 4  
##   xbar    SD      n T_obs  
##   <dbl> <dbl> <int> <dbl>  
## 1  237.  44.6  4190  9.73
```

One Sample Hypothesis Test: Example II



So, is 9.725 far from 0?

IF the null hypothesis is true, what is the probability of being further away?
The area outside of:



One Sample Hypothesis Test: Example II



Find the p-value:

$$\begin{aligned} P(\text{more extreme} \mid H_0 \text{ true}) &= P(T_{4189} > 9.725) + P(T_{4189} < -9.725) \\ &= 2 * P(T_{4189} < -9.725) \end{aligned}$$

Very small:

```
T_4189 <- StudentsT(df = 4189)
2*cdf(T_4189, -9.725)
```

```
## [1] 4.036102e-22
```

Since $p < \alpha$, the probability of something "more extreme" is small, so our data is very extreme.

Our data are far from the null hypothesis, so why would we still think the null hypothesis is true? We reject the null.

One Sample Hypothesis Test: Example II



This is not very constructive: sure, the true mean is not 230 mg/dL... but what is it then?

Let's try to characterize all the values of μ_0 that we would NOT reject. I.e. which $H_0 : \mu = \mu_0$ would not be rejected?

If \bar{X} is close enough to μ_0 , then we wouldn't not reject the null.

"Close enough" means that the probability of being "more extreme" is greater than $\alpha = 0.05$. In other words, $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ has to be so close to zero that

$$\begin{aligned} P(\text{more extreme} \mid H_0) &= P(T_{n-1} > |T_{\text{obs}}|) + P(T_{n-1} < -|T_{\text{obs}}|) \\ &= 2 \cdot P(T_{n-1} > |T_{\text{obs}}|) < 0.05. \end{aligned}$$

I.e. $P(T_{n-1} > |T_{\text{obs}}|) = 0.025$.

One Sample Hypothesis Test: Example II



That means that if $T_{\text{obs}} > t_{n-1,0.025}$ or $T_{\text{obs}} < -t_{n-1,0.025}$, then we would reject. Or, in other words, if $-t_{n-1,0.025} < T_{\text{obs}} < t_{n-1,0.025}$, then we would not reject. So, we do NOT reject when

$$-t_{n-1,0.025} < \frac{\bar{X} - \mu_0}{s/\sqrt{n}} < t_{n-1,0.025}$$

or equivalently, when

$$\bar{X} - t_{n-1,0.025} \frac{s}{\sqrt{n}} < \mu_0 < \bar{X} + t_{n-1,0.025} \frac{s}{\sqrt{n}}$$

This is **EXACTLY** the 95% confidence interval we would calculate for μ !!

One Sample Hypothesis Test: Example II



So,

reject $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ at significance level α



μ_0 is NOT in the $(1 - \alpha) \cdot 100\%$ CI for μ

This supports our interpretation of CIs as a "collection of plausible values", i.e. values we wouldn't reject based on the data we have.

One Sample Hypothesis Test: Example II



For the total cholesterol, let's find a 95% confidence interval:

```
T_4189 <- StudentsT(df = 4189)

fram %>%
  summarize(xbar = mean(totChol),
            SD = sd(totChol),
            n = n(),
            LL = xbar - quantile(T_4189, 0.975)*SD/sqrt(n),
            UL = xbar + quantile(T_4189, 0.975)*SD/sqrt(n))

## # A tibble: 1 x 5
##   xbar    SD      n    LL    UL
##   <dbl> <dbl> <int> <dbl> <dbl>
## 1  237.  44.6  4190  235.  238.
```

As expected, the value we just rejected (230 mg/dL) is NOT in the 95% confidence interval.

One Sample Hypothesis Test: Example I (revisited)



Let's revisit the paint thickness data and calculate a 95% confidence interval:

```
paint_thickness_CI <- paint_thickness %>%  
  summarize(xbar = mean(thickness),  
            SD = sd(thickness),  
            n = n(),  
            LL = xbar - quantile(StudentsT(df = n - 1), 0.975)*SD/sqr  
            UL = xbar + quantile(StudentsT(df = n - 1), 0.975)*SD/sqr  
  
paint_thickness_CI
```

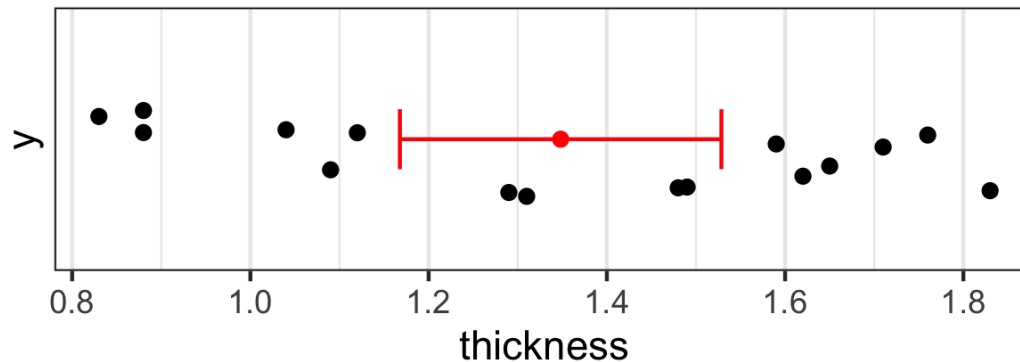
```
## # A tibble: 1 x 5  
##   xbar    SD    n    LL    UL  
##   <dbl> <dbl> <int> <dbl> <dbl>  
## 1  1.35 0.339   16  1.17  1.53
```

We see that the value we did not reject, i.e. 1.5, is indeed in the 95% confidence interval, as we would expect.

One Sample Hypothesis Test: Example I (revisited)



```
ggplot(data = paint_thickness,  
       aes(x = thickness, y = 1)) +  
  geom_jitter(width = 0, height = 0.1) +  
  ## Below we add some things to the plot using a different data set,  
  ## namely paint_thickness_CI, which has the things we just calculated  
  geom_errorbarh(data = paint_thickness_CI,  
                 aes(xmin = LL, xmax = UL, x = NULL),  
                 height = 0.1, color = "red") +  
  geom_point(data = paint_thickness_CI,  
             aes(x = xbar), color = "red") +  
  scale_y_continuous(limits = c(0.8, 1.2), breaks = NULL, minor_breaks = NULL)
```



One Sample Hypothesis Test: Example III



```
library(lubridate)
corona_virus <- read_csv(here::here("csv_data/corona_virus.csv")) %>%
  mutate(date = mdy(date))

corona_virus %>% DT::datatable(options = list(pageLength = 5, scroll))
```

Show entries

Search:

	Province/State	Country/Region	Lat	Long	date	Confirmed	Deaths
1	Anhui	Mainland China	31.8257	117.2264	2020-01-22	1	0
2	Anhui	Mainland China	31.8257	117.2264	2020-01-23	9	0
3	Anhui	Mainland China	31.8257	117.2264	2020-01-24	15	0
4	Anhui	Mainland China	31.8257	117.2264	2020-01-25	39	0
5	Anhui	Mainland China	31.8257	117.2264	2020-01-26	60	0

Showing 1 to 5 of 6,880 entries

Previous

1

2

3

4

5

...

1376

Next

One Sample Hypothesis Test: Example III



Using this data, we can estimate the risk of dying:

```
corona_virus %>%  
  filter(date == max(date)) %>%  
  summarize_at(.vars = vars(Confirmed, Deaths, Recovered),  
               .funs = sum) %>%  
  mutate(P = Deaths/Confirmed)
```

```
## # A tibble: 1 x 4  
##   Confirmed Deaths Recovered      P  
##   <dbl>   <dbl>   <dbl>  <dbl>  
## 1    95124    3254    51171 0.0342
```

One Sample Hypothesis Test: Example III



Obviously many problems with this:

- data highly dependent
- would definitely need a time component
- ...

For now, let's ignore this.

One Sample Hypothesis Test: Example III



Researchers might be interested in knowing if the risk of dying from the corona virus is dependent on region. They want to test if the risk in Italy is the same as the risk in Mainland China.

NOTE: the following is not how you would actually do this, but is the best we can do with our current tools. We will revisit this later.

First, they estimate the risk in Mainland China and Italy:

```
corona_virus %>%  
  filter(`Country/Region` %in% c("Italy", "Mainland China"),  
         date == max(date)) %>%  
  group_by(`Country/Region`) %>%  
  summarize_at(vars(Confirmed, Deaths, Recovered),  
               sum) %>%  
  mutate(P = Deaths/Confirmed)
```

```
## # A tibble: 2 x 5  
##   `Country/Region` Confirmed Deaths Recovered      P  
##   <chr>           <dbl>   <dbl>     <dbl>  <dbl>  
## 1 Italy           3089     107       276 0.0346  
## 2 Mainland China 80271    2981     49955 0.0371
```

One Sample Hypothesis Test: Example III



If we take the estimated risk in Mainland China as "the true risk" in Mainland China, we can test $H_0 : \pi_{\text{Italy}} = 0.0371$ vs. $H_A : \pi_{\text{Italy}} \neq 0.0371$.

We will use $\alpha = 0.1$, and estimate π_{Italy} using $P = \frac{\# \text{deaths}}{\# \text{confirmed}} = \frac{1}{n} \sum_{i=1}^n X_i$, where $X_i \sim \text{Bernoulli}(\pi_{\text{Italy}})$.

As always, we want to see if our estimate P is far from our null 0.0371, and we still want to compare it to a measure of variation of our estimator P . So think about using

$$\frac{P - 0.0371}{\text{SD}(P)}$$

IF the null hypothesis is true, then $0.0371 = E(P)$.

Remember, $\text{SD} = \sqrt{\pi_{\text{Italy}}(1 - \pi_{\text{Italy}})/n}$. **IF** the null hypothesis is true, then $\pi_{\text{Italy}} = 0.0371$, so $\text{SD}(P) = \sqrt{0.0371 \cdot (1 - 0.0371)/n}$.

So **IF** the null hypothesis is true, then

One Sample Hypothesis Test: Example III



Now, remember that $P \sim N$, if $n \cdot \pi_{\text{Italy}} > 5$ and $n(1 - \pi_{\text{Italy}}) > 5$.

IF the null is true, then $P \sim N$ if $n \cdot 0.0371 > 5$ and $n \cdot (1 - 0.0371) > 5$.

n is the number of confirmed in Italy, so $n = 3089$. Since $3089 \cdot 0.0371 = 114.6019$ and $3089 \cdot (1 - 0.0371) = 2974.3981$, the CLT tells us that $P \sim N$, **IF** H_0 is true.

So, **IF** H_0 is true, then

$$Z = \frac{P - E(X)}{\text{SD}(P)} = \frac{P - 0.0371}{\sqrt{\frac{0.0371(1-0.0371)}{3089}}} \sim N(0, 1).$$

One Sample Hypothesis Test: Example III



So, to test the null hypothesis $H_0 : \pi_{\text{Italy}} = 0.0371$ vs. $H_A : \pi_{\text{Italy}} \neq 0.0371$, we need to find the probability of something *more extreme*.

First, we find Z_{obs} :

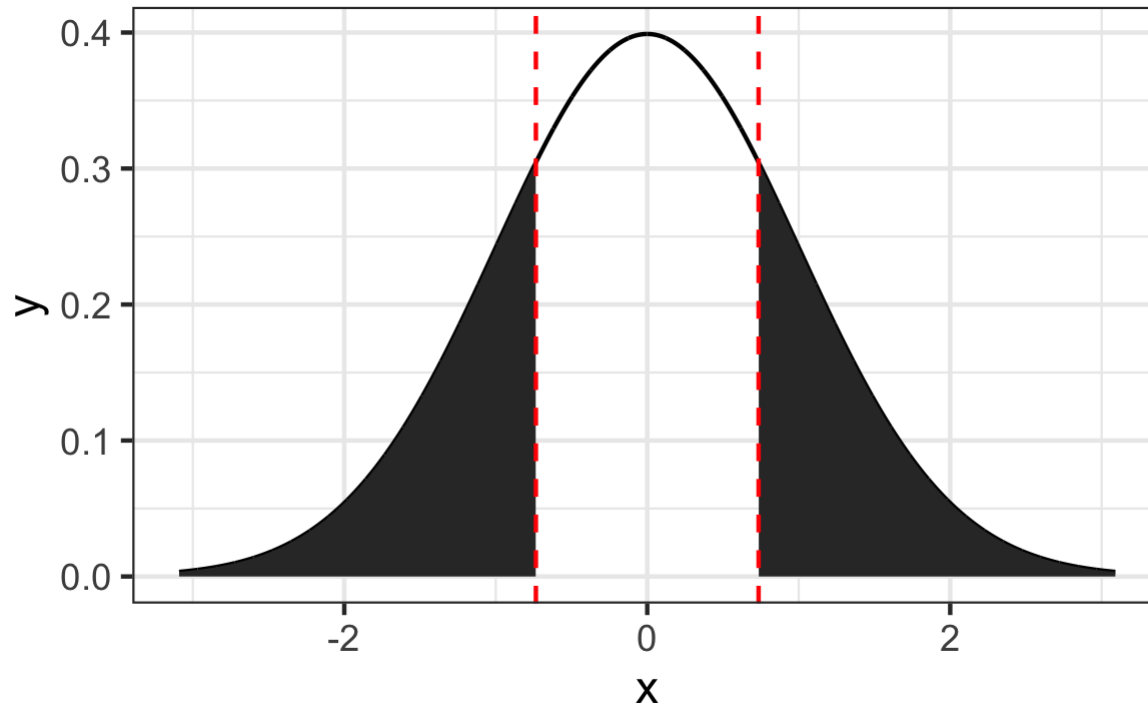
$$Z_{\text{obs}} = \frac{\hat{p} - 0.0371}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{0.0346 - 0.0371}{\sqrt{\frac{0.0371(1-0.0371)}{3089}}} \approx -0.7351$$

Then we want to find $P(\text{more extreme} | H_0 \text{ true})$. What is more extreme in this case? Smaller than -0.7351 or greater than 0.7351.

One Sample Hypothesis Test: Example III



So, p-value = $P(Z < -0.7351) + P(Z > 0.7351) = 2 \cdot P(Z < -0.7351)$.



```
Z <- Normal(mu = 0, sigma = 1)
```

```
2*cdf(Z, -0.7351)
```

```
## [1] 0.4622786
```

Since this is much larger than our $\alpha = 0.1$, we do not reject the idea that the mortality rates are the same in Italy and Mainland China.

One Sample Hypothesis Test: Example III



90% CIs for the risk in Mainland China and Italy:

```
corona_virus %>%  
  filter(`Country/Region` %in% c("Italy", "Mainland China"),  
         date == max(date)) %>%  
  group_by(`Country/Region`) %>%  
  summarize_at(.vars = vars(Confirmed, Deaths, Recovered),  
               .funs = sum) %>%  
  mutate(P = Deaths/Confirmed,  
         LL = P - quantile(Normal(), 0.95)*sqrt(P*(1-P)/Confirmed),  
         UL = P + quantile(Normal(), 0.95)*sqrt(P*(1-P)/Confirmed))
```

```
## # A tibble: 2 x 7  
##   `Country/Region` Confirmed Deaths Recovered      P      LL      UL  
##   <chr>           <dbl>   <dbl>    <dbl> <dbl> <dbl> <dbl>  
## 1 Italy           3089     107      276 0.0346 0.0292 0.0401  
## 2 Mainland China 80271    2981    49955 0.0371 0.0360 0.0382
```

One Sample Hypothesis Test: Example III



What assumptions did we make that most likely are NOT true?

- Outcome of each confirmed case is a Bernoulli RV with some probability of success
- All independent

Other problems:

- When comparing rate in Italy to rate in Mainland China, we ignored that the rate in Mainland China was actually *estimated* from the data.

One Sample Hypothesis Test: Example III



A map:

```
world_map <- map_data("world")

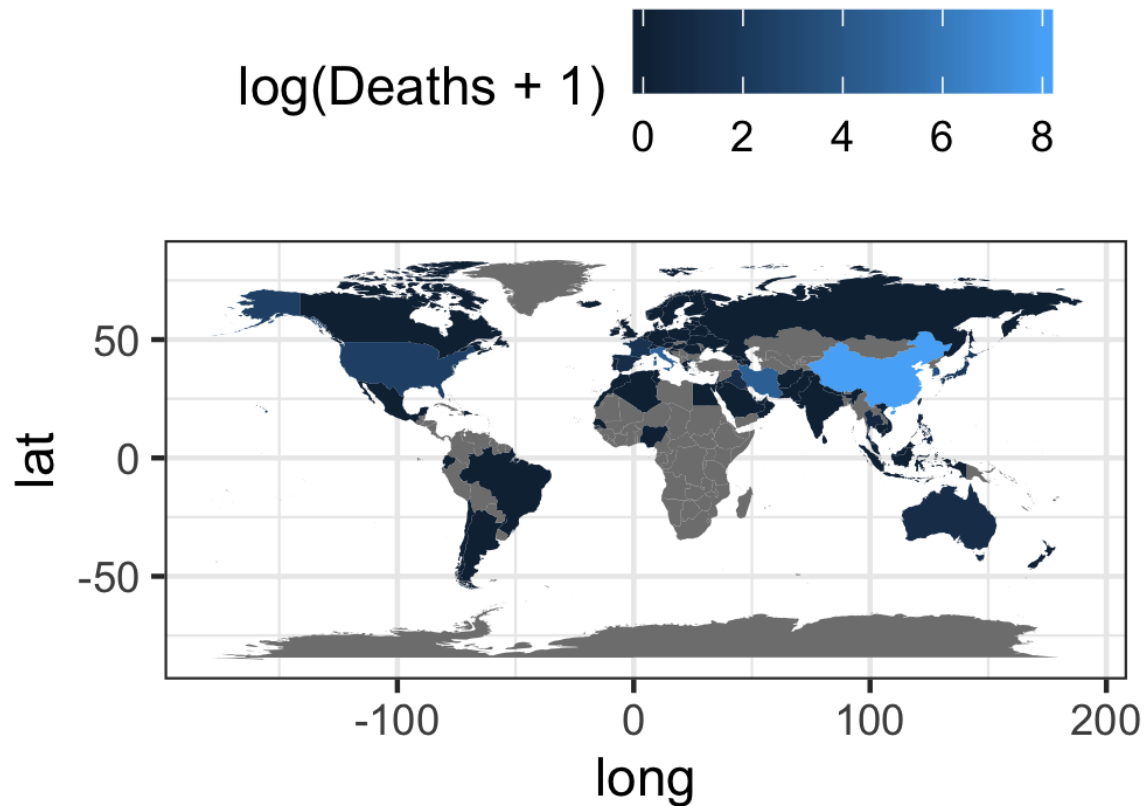
for_map <- corona_virus %>%
  filter(date == max(date)) %>%
  group_by(region = `Country/Region`) %>%
  summarize_at(vars(Confirmed, Deaths, Recovered), sum) %>%
  ungroup() %>%
  mutate(region = case_when(region == "Mainland China" ~ "China",
                             region == "US" ~ "USA",
                             TRUE ~ region))

ggplot() +
  geom_map(data = left_join(world_map, for_map),
           map = world_map,
           aes(x = long, y = lat, map_id = region,
               fill = log(Deaths+1))) +
  coord_fixed() +
  theme(legend.position = "top")
```

One Sample Hypothesis Test: Example III



A map:



One Sample Hypothesis Test: Example III



More maps:

```
for_map2 <- for_map %>%  
  pivot_longer(cols = 2:4, names_to = "type", values_to = "counts") %  
  filter(!is.na(counts), !is.na(type)) %>%  
  group_by(region) %>%  
  nest()  
  
for_map3 <- world_map %>%  
  select(region) %>%  
  unique() %>%  
  left_join(for_map2) %>%  
  mutate(data = map(data,  
    function(x){  
      if(length(x) == 0){  
        tibble(type = c("Confirmed", "Deaths", "Recovered"),  
          counts = rep(NA, 3))  
      } else { x }  
    }) %>%  
  unnest(cols = data)  
  
ggplot() +  
  geom_map(data = left_join(world_map, for_map3),  
    map = world_map,  
    aes(x = long, y = lat, map_id = region, fill = log(counts+1)),  
    coord_fixed() + facet_wrap(type ~ ., ncol = 2) + theme(legend.position = "right")
```

One Sample Hypothesis Test: Example III



More maps: