Lecture 13: One Sample Hypothesis Tests

STAT 324

Ralph Trane

University of Wisconsin–Madison

Spring 2020

One Sample Hypothesis Test: Example IV

- Secondhand smoke is of great health concern, especially for children.
- A sample was taken of 15 children in foster care suspected of being exposed to secondhand smoke, and the amount of cotinine (a metabolite of nicotine) in the urine was measured in ng/mL.
- Cotinine in unexposed children should be below 75 units. The data were as follows:

```r
library(tidyverse); theme_set(theme_bw())
secondhand_smoking <- tibble(cotinine = c( 29,  30,  53,  75,  34,
                                           21,  12,  58, 117, 119,
                                          115, 134, 253, 289, 287))
```

Based on this data, does it seem like the mean of cotinine of children exposed to secondhand smoke is greater than 75 ng/mL?

One Sample Hypothesis Test: Example IV

We would like to test $H_0 : \mu = 75$ vs. $H_A : \mu > 75$.

What significance level would you pick?

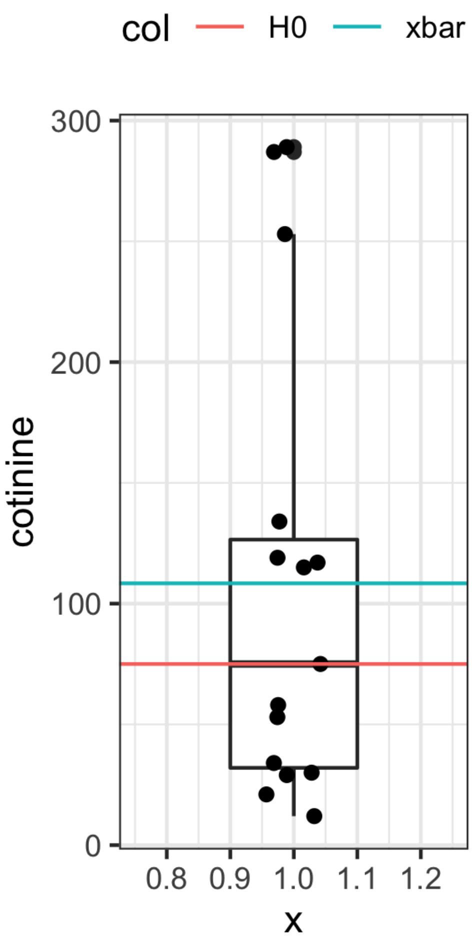
Recall, $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{true})$. Also, decreasing $P(\text{type I error})$ increases $P(\text{type II error})$ (will see why later). So, picking α can be done by assessing which is worse: type I or type II error.

Here, I would say type II error is worse: if we fail to reject the null, but we should have, we conclude that the children are fine, when in reality the levels of cotinine is too high.

So, let's pick $\alpha = 0.1$, which is traditionally relatively "large".

One Sample Hypothesis Test: Example IV

```
hlines <- tibble(
  y = c(mean(secondhand_smoking$
        75),
  col = c('xbar', 'H0')
)

ggplot(secondhand_smoking,
       aes(x = 1, y = cotinine))
  geom_boxplot(width = 0.2) +
  geom_jitter(height = 0,
              width = 0.05) +
  geom_hline(data = hlines,
             aes(yintercept = y,
                 color = col)) +
  xlim(c(0.75, 1.25)) +
  theme_bw() +
  theme(legend.position = "top")
```



One Sample Hypothesis Test: Example IV

Since the mean is our parameter of interest, we would like to use $T = \frac{\bar{X} - 75}{\widehat{\text{SD}}(\bar{X})}$ as our metric for being "far from the null".

Here, since $H_A : \mu > 75$, we would reject H_0 if T is much larger than 0.

First, we calculate $T_{\text{obs}} = \frac{\bar{x}_{\text{obs}} - 75}{\widehat{\text{SD}}(\bar{X})}$. Per usual, we will use $\widehat{\text{SD}}(\bar{X}) = \hat{\sigma}/\sqrt{n} = S/\sqrt{n}$.

```
sum_stats <- secondhand_smoking %>%
  summarize(xbar = mean(cotinine),
            SD = sd(cotinine),
            n = n(),
            T_obs = (xbar - 75)/(SD/sqrt(n)))
sum_stats
```

```
## # A tibble: 1 x 4
##    xbar    SD     n T_obs
##   <dbl> <dbl> <int> <dbl>
## 1  108.  95.6    15  1.35
```

One Sample Hypothesis Test: Example IV

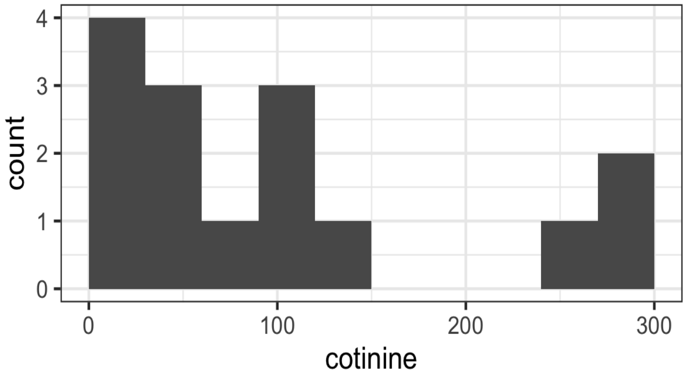
To assess if 1.3531105 is "much larger than 0", we want to calculate $P(T > 1.3531105 \mid H_0 \text{ true})$.

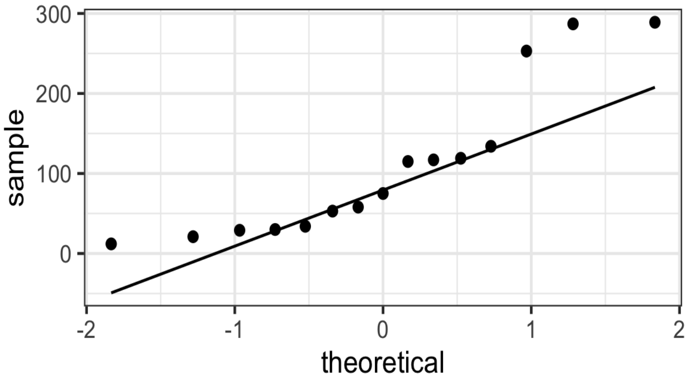
If $\bar{X} \sim N$, then $T \sim t_{n-1}$. So, need to check if $\bar{X} \sim N$.

Small n, so CLT might be a stretch. Let's consider the data:

```
ggplot(data = secondhand_smoking
       aes(x = cotinine)) +
  geom_histogram(binwidth = 30,
```

```
ggplot(data = secondhand_smoking
       aes(sample = cotinine)) +
  geom_qq() + geom_qq_line()
```





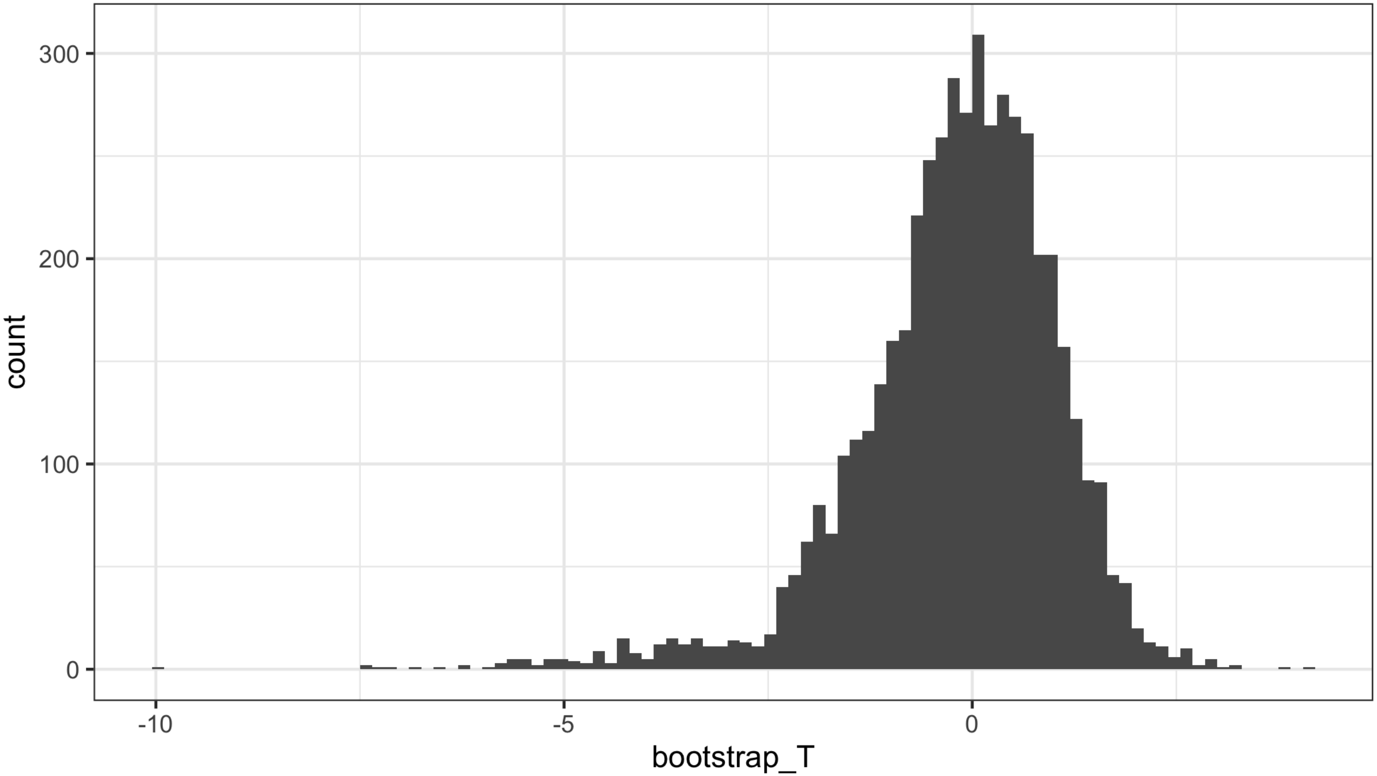
One Sample Hypothesis Test: Example IV

So it does NOT seem like $\bar{X} \sim N$. What else can we do to find the distribution of $T = \frac{\bar{X} - 75}{\widehat{\text{SD}}(\bar{X})}$?

Bootstrap!

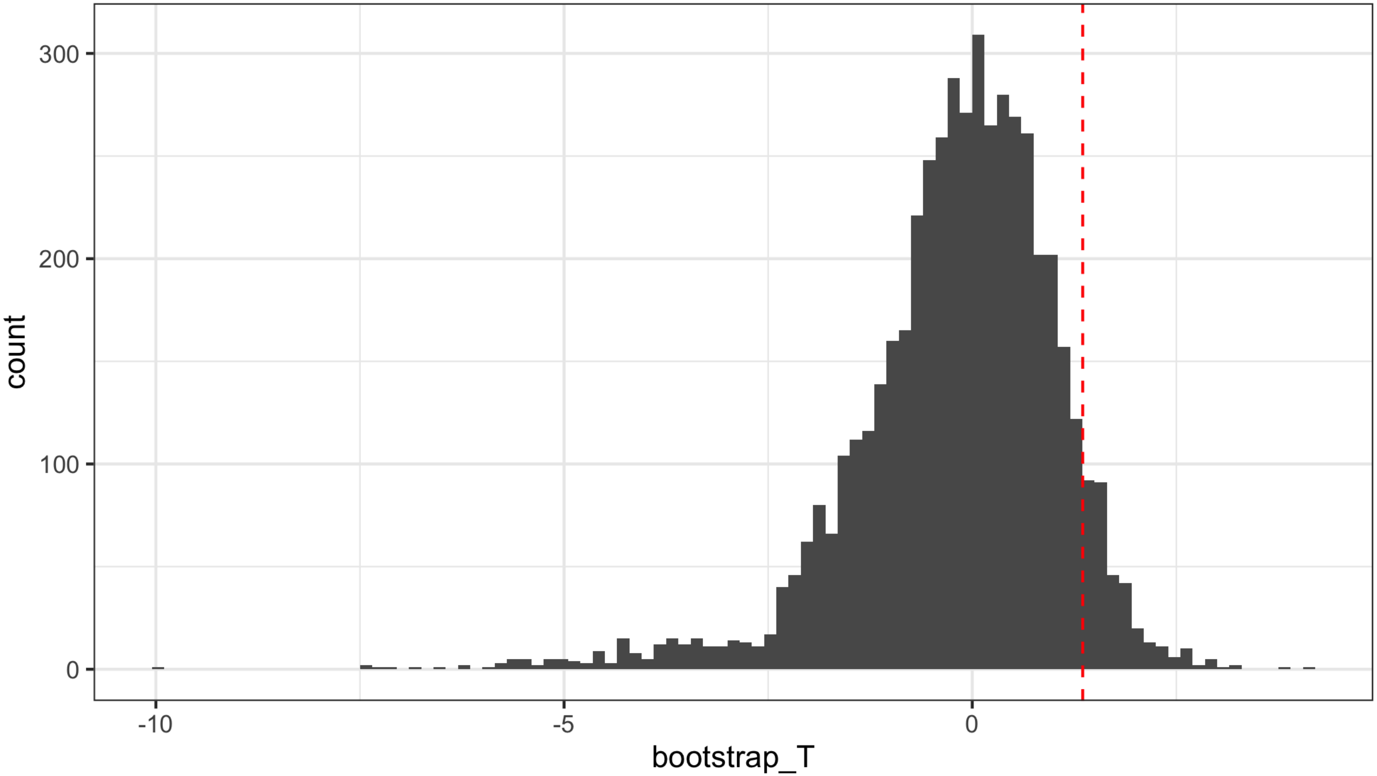
```
xbar_orig <- sum_stats$xbar

bootstrap_samples <- tibble(i = 1:5000) %>%
  mutate(bootstrap_sample = map(i,
                                ~sample_n(secondhand_smoking,
                                          size = 15,
                                          replace = TRUE)$cotinine),
         bootstrap_mean = map_dbl(bootstrap_sample, mean),
         bootstrap_sd = map_dbl(bootstrap_sample, sd),
         bootstrap_T = (bootstrap_mean - xbar_orig)/
           (bootstrap_sd/sqrt(15)))

ggplot(bootstrap_samples,
       aes(x = bootstrap_T)) +
  geom_histogram(binwidth = 0.15, boundary = 0)
```

One Sample Hypothesis Test: Example IV



Now that we know the distribution of T, we can find $P(T > 1.3531105)$. But how?

One Sample Hypothesis Test: Example IV



Now that we know the distribution of T, we can find $P(T > 1.3531105)$. But how? Proportion of the area under the curve to the right of 1.3531105.

One Sample Hypothesis Test: Example IV

The proportion of the area under the curve to the right of 1.3531105 is the same as the proportion of bootstrap T's that are greater than 1.3531105. So $\text{p-value} = P(T > 1.3531105) = \frac{\# T_{\text{boots}} > 1.3531105}{5000}$.

```
n_larger <- sum(bootstrap_samples$bootstrap_T > 1.3531105)
n_larger
```

```
## [1] 342
```

```
p_value <- n_larger/5000
p_value
```

```
## [1] 0.0684
```

Since the p-value is smaller than our significance level $\alpha = 0.1$, we reject the null hypothesis: these data suggest that the true mean amount of cotinine in the urine is greater than 75 ng/mL.

One Sample Hypothesis Tests: summary

One Sample T-test

If $\bar{X} \sim N$, the hypothesis $H_0 : \mu = \mu_0$ can be tested using the *test-statistic* $T = \frac{\bar{X} - \mu_0}{\widehat{\mathrm{SD}}(\bar{X})}$.

Under the null hypothesis, $T \sim t_{n-1}$, and we reject H_0 if the $\text{p-value} = P(T \text{ more extreme than } T_{\text{obs}} \,|H_0) < \alpha$.

- Here, $\widehat{\mathrm{SD}}(\bar{X}) = s/\sqrt{n}$

One Sample Test for Proportion

If $P \sim N$ (approximately), the hypothesis $H_0 : \pi = \pi_0$ can be tested using the *test-statistic* $Z = \frac{P - \pi_0}{\mathrm{SD}_0(P)}$.

Under the null hypothesis, $Z \sim N(0, 1)$, and we reject H_0 if the $\text{p-value} = P(Z \text{ more extreme than } Z_{\text{obs}} \,|H_0) < \alpha$.

- Here, $\mathrm{SD}_0(P) = \sqrt{\pi_0(1 - \pi_0)/n}$ is the standard deviation of P **IF** the null hypothesis is true.

One Sample Hypothesis Tests: summary

One Sample Bootstrap Test

If $\bar{X}$ is NOT normally distributed, but the observations are independent, we can test $H_0 : \mu = \mu_0$ using the *test-statistic* $T = \frac{\bar{X}-\mu_0}{\widehat{\text{SD}}(\bar{X})}$.

We estimate the distribution of T under the null by the distribution of T_{boot}, and reject H_0 if the p-value = $P(T_{\text{boot}} \text{ more extreme than } T_{\text{obs}}) < \alpha$.

Here,

- $B =$ number of bootstrap samples created.
- $T_{\text{boot}} = \frac{\bar{x}_{\text{boot}}-\bar{x}_{\text{obs}}}{s_{\text{boot}}/\sqrt{n}}$ is calculated for each of the bootstrap samples.

One Sample Hypothesis Tests: summary

"More extreme"

What it means to be more extreme is determined by the alternative hypothesis:

- if $H_A : \mu > \mu_0$, then "more extreme" = "greater than", and p-value = $P(T > T_{\text{obs}})$
- if $H_A : \mu < \mu_0$, then "more extreme" = "smaller than", and p-value = $P(T < T_{\text{obs}})$
- if $H_A : \mu \neq \mu_0$, then "more extreme" = "further from 0", and p-value = $P(T < -|T_{\text{obs}}|) + P(T > |T_{\text{obs}}|)$

For test of proportion, replace μ with π, and T with Z.

One Sample Hypothesis Test: Power

Recall: for every hypothesis test, there is a conclusion. For every conclusion, one of three things will happen:

- we make the right decision
  - i.e. reject when H_0 is false, do not reject when H_0 is true
- we make a type I error
  - i.e. reject when H_0 is actually true
- we make a type II error
  - i.e. fail to reject when H_0 is actually false

One Sample Hypothesis Test: Power

We have full control over the type I error rate - it is exactly α. To see this, say we are testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$.

$$\begin{aligned} P(\text{reject } H_0 \mid H_0 \text{ true}) &= P(T \text{ very far from } 0 \mid H_0 \text{ true}) \\ &= P(T > t_{n-1,\alpha/2} \mid H_0 \text{ true}) + P(T < -t_{n-1,\alpha/2} \mid H_0 \text{ true}) \\ &= \alpha \end{aligned}$$

Similarly can be done for the two other alternative hypotheses.

One Sample Hypothesis Test: Power

The Type II error rate is a bit more tricky:
$P(\text{type II}) = P(\text{fail to reject } H_0 \mid H_0 \text{ true})$.

To get a better sense of how this works, we will consider a simple case that we haven't talked about yet: one sample hypothesis test with known variance σ^2.

Remember: good old paint thickness data.

```
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)
)
```

Let us assume that $\bar{X} \sim N$, and that we know the true variance $\sigma^2 = 0.115$.

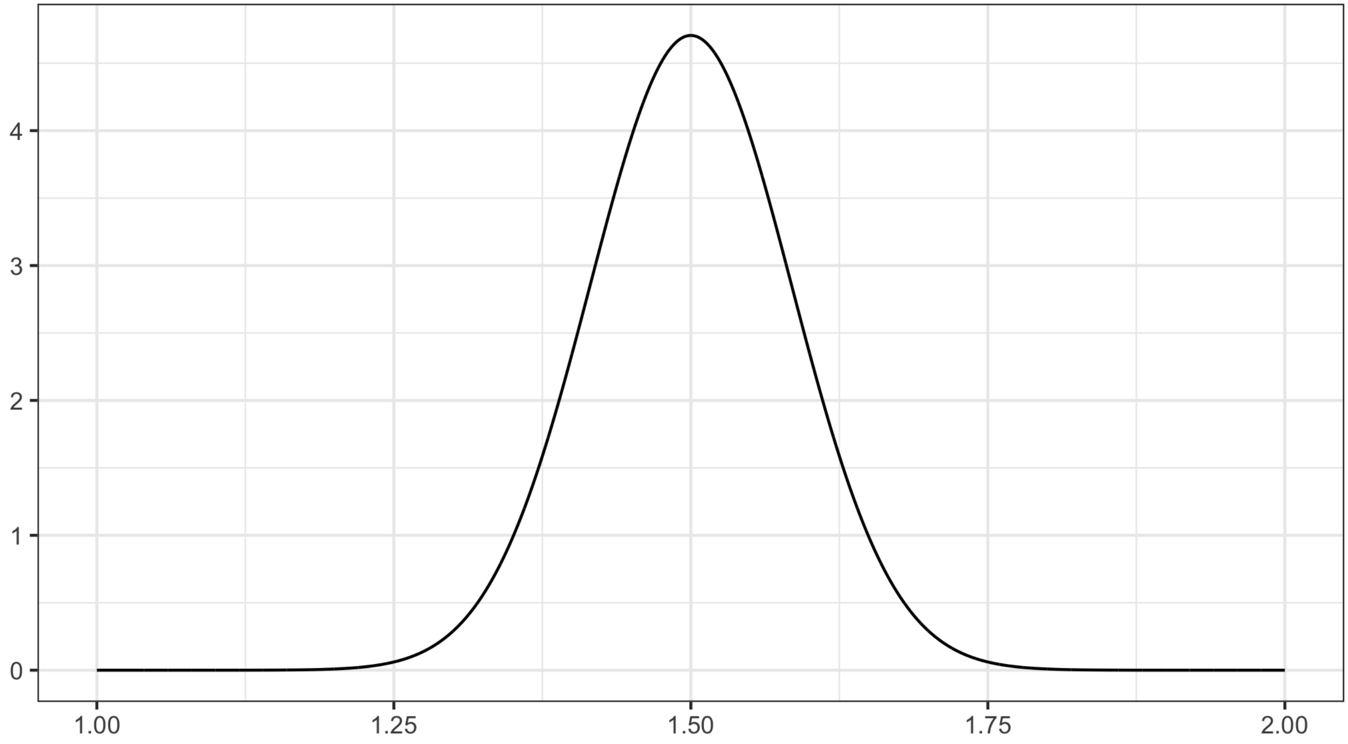
One Sample Hypothesis Test: Power

When testing $H_0 : \mu = 1.5$ against $H_A : \mu \neq 1.5$ using $\alpha = 0.05$, we would reject when $\bar{x}_{\text{obs}}$ is very, very far from 1.5.

Actually, we reject when $\bar{x}_{\text{obs}}$ is so far from 1.5 that the probability of $\bar{X}$ being even further from 1.5 is less than 0.05: $2 \cdot P(\bar{X} > |\bar{x}_{\text{obs}}| \mid H_0 \text{ true}) < 0.05$.

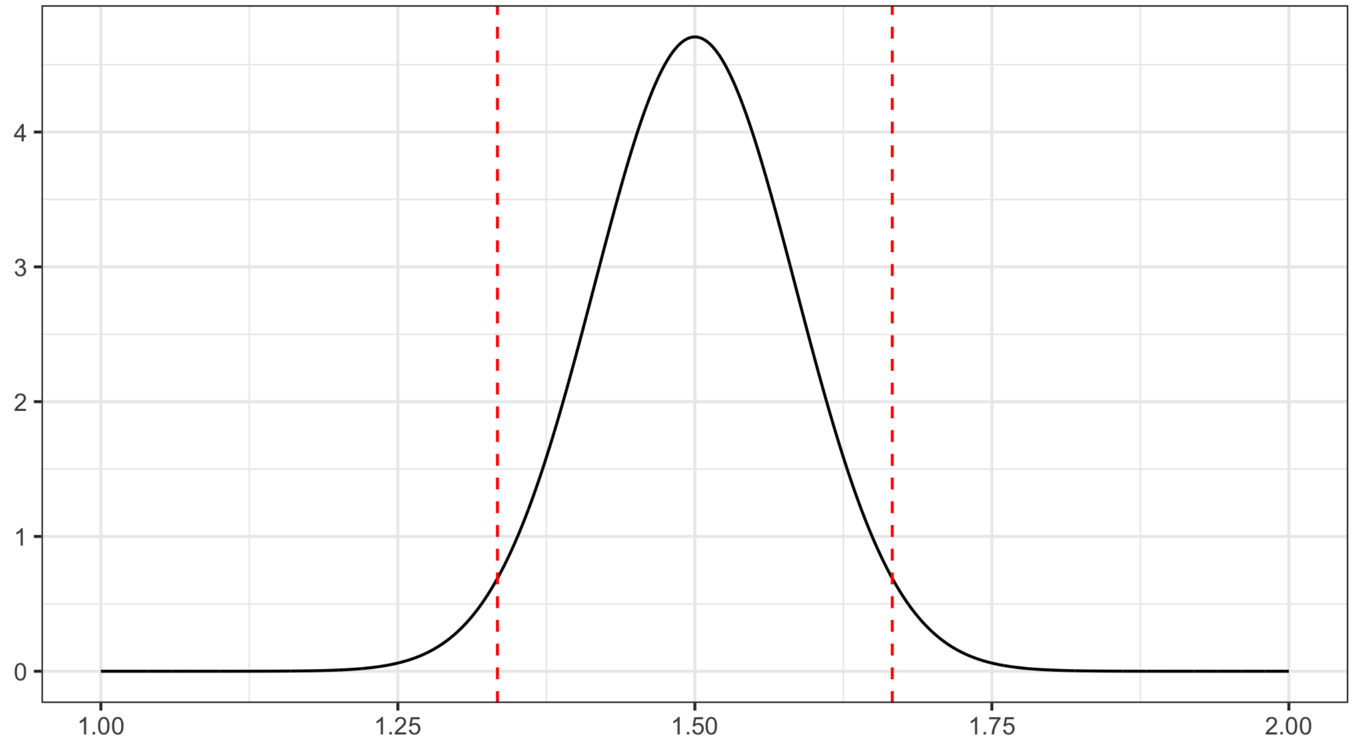
One Sample Hypothesis Test: Power

When H_0 is true, $\bar{X} \sim N(1.5, 0.115/16)$. So, when H_0 is true, $\bar{X}$ follows this distribution:



One Sample Hypothesis Test: Power

Will reject when $\bar{X}$ outside red dotted lines.



But to talk about type II error rate (or power), we need to find
$P(\text{fail to reject } H_0 \mid H_0 \text{ not true}) = P(\text{fail to reject } H_0 \mid H_A \text{ true})$ (or

One Sample Hypothesis Test: Power

Assuming we know the true $\sigma^2 = 0.115$, and that the null hypothesis is true, then $\bar{X} \sim N(1.5, 0.115/16)$.

So, we reject when $\bar{X}$ outside red dotted lines. I.e. if $0.025 > P\left(\bar{X} > |\bar{x}_{\text{obs}}|\right)$.

```r
X_H0 <- Normal(mu = 1.5, sigma = sqrt(0.115/16))

quantile(X_H0, c(0.025, 0.975))
```

```
## [1] 1.333836 1.666164
```

So, we reject when $\bar{x} < 1.33$ or $\bar{x} > 1.66$.

One Sample Hypothesis Test: Power

But what if the true mean is NOT 1.5? What if, instead, it is 1.3? What is $P(\text{type II error})$?

To find the probability of rejecting, we need to look at a different curve, because if $\mu = 1.3$, then $\bar{X} \sim N(1.3, 0.115/16)$.

One Sample Hypothesis Test: Power

But what if the true mean is NOT 1.5? What if, instead, it is 1.3? What is $P(\text{type II error})$?

To find the probability of rejecting, we need to look at a different curve, because if $\mu = 1.3$, then $\bar{X} \sim N(1.3, 0.115/16)$.

$$\begin{aligned} P(\text{type II error}) &= P(\text{fail to reject} \mid \mu = 1.3) \\ &= P(1.33 < \bar{X} < 1.66 \mid \mu = 1.3) \\ &= P(\bar{X} < 1.66 \mid \mu = 1.3) - P(\bar{X} < 1.33 \mid \mu = 1.3) \end{aligned}$$

```
X_HA <- Normal(1.3, sqrt(0.115/16))

cdf(X_HA, 1.66) - cdf(X_HA, 1.33)
```

```
## [1] 0.3617108
```

That is, **IF** the true mean is 1.3, we would fail to reject the idea that $\mu = 1.5$ about 37% of the time.

Or, similarly, we would only reject $\mu = 1.5$ about 63% of the time.

One Sample Hypothesis Test: Power

What sample size do we need to be able to distinguish between $\mu = 1.5$ and $\mu = 1.3$ most of the time? Say, 80% of the time?

That is, what should n be such that $P(\text{reject } H_0 : \mu = 1.5 \mid \mu = 1.3) = 0.8$?

More general, if we are testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ at significance level α, what sample size is needed to make sure that $P(\text{reject } H_0 \mid \mu = \mu_A) = 1 - \beta$?

Turns out, this is approximately

$$n \approx \left(\frac{\sigma(z_{\alpha/2} + z_{\beta})}{\mu_0 - \mu_A} \right)^2$$

In our specific example, $\sigma = 0.34$, $\alpha = 0.05$, $\beta = 0.2$, $\mu_0 = 1.5$, and $\mu_A = 1.3$.

One Sample Hypothesis Test: Power

We can find $z_{0.025}$ and z_β:

```
Z <- Normal()
quantile(Z, c(1-0.025, 1-0.2))
```

```
## [1] 1.9599640 0.8416212
```

So, $n \approx \left(\frac{0.34(1.96+0.84)}{1.5-1.3}\right)^2 = 22.6576.$

To have 80% power to reject 1.5 as the true mean if the true mean is in fact 1.3, we would need 23 samples.

One Sample Hypothesis Test: Power

- sample size depends only on standard deviation, α, β, and *difference* between true and hypothesized means.

- in practice, it goes as follows:

  - researchers pick desired α and β
  - from previous, well-done experiments, a solid estimate of σ is optained
  - from expert knowledge, a "minimal interesting difference" is chosen (i.e. $\mu_0 - \mu_A$)
  - based on this, needed sample size is determined.