Lecture 16: Two Sample Hypothesis Tests

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Two Independent Samples Hypothesis Test

Last time

Two sample T-tests. Objective: test the hypothesis $H_0 : \mu_1 - \mu_2 = v_0$ against an alternative. (We only considered $v_0 = 0$, but could be any number. We only considered $H_A : \mu_1 - \mu_2 \neq v_0$, but could be any of the three $>, <, \neq$.)

Assumptions:

- the two samples are independent of each other
- the observations in each sample are independent
- and the averages $\bar{X}_1$ and $\bar{X}_2$ are normally distributed.

Test statistic: $T = \frac{V - v_0}{\widehat{\mathrm{SD}}(V)}$ which follows a t-distribution **IF** the null hypothesis is true.

Two Independent Samples Hypothesis Test

Last time

Two scenarios determines how we calculate $\widehat{\mathrm{SD}}(V)$, and the degrees of freedom for the t-distribution:

If $0.5 < \frac{s_1}{s_2} < 2$, we assume equal variances $\sigma_1^2 = \sigma_2^2$. In this case,

- $\widehat{\mathrm{SD}}(V) = s_p\sqrt{1/n_1 + 1/n_2}$, where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- if H_0 is true, $T \sim t_{n_1+n_2-2}$.

If we cannot assume equal variances,

- $\widehat{\mathrm{SD}}(V) = \sqrt{s_1^2/n_1 + s_2^2/n_2}$
- if H_0 is true, $T \sim t_\nu$ where

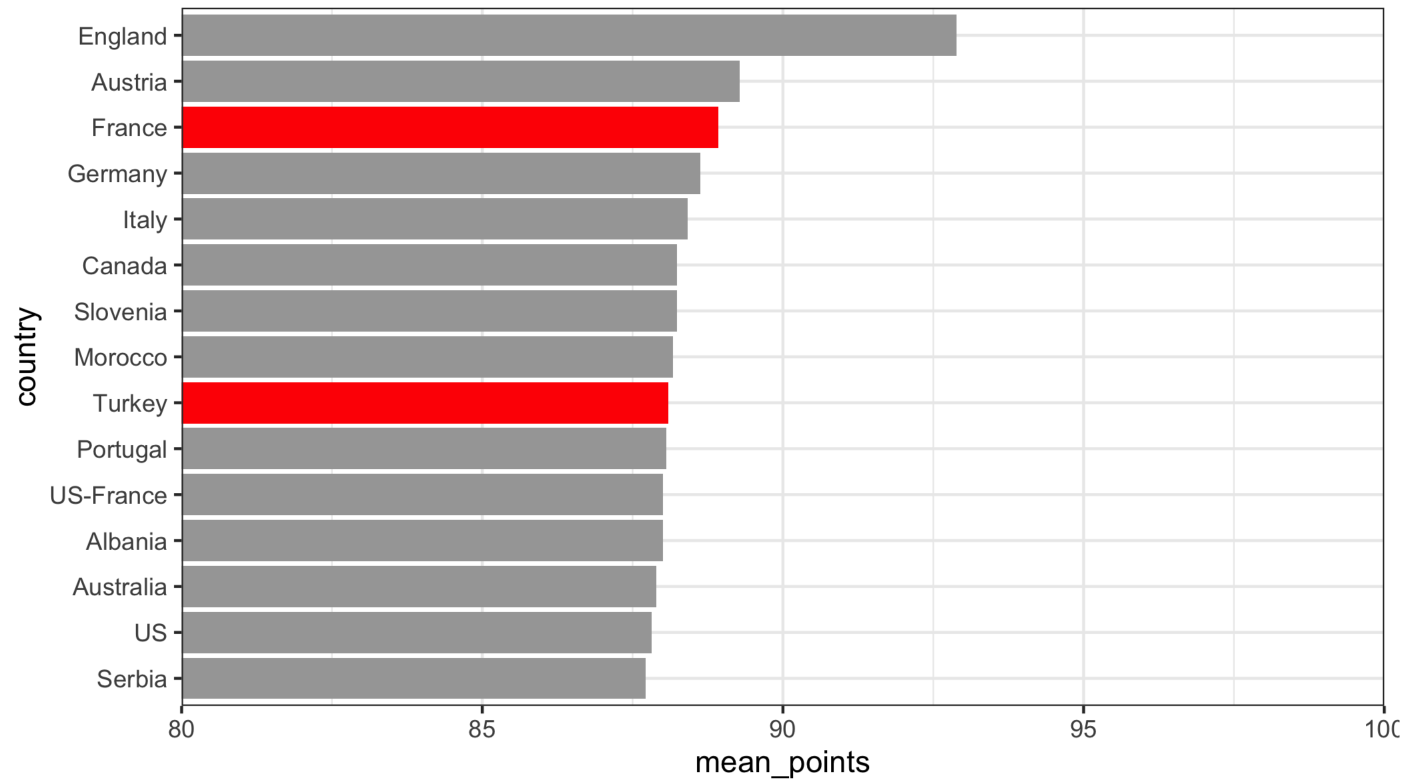
$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

Two Independent Samples Hypothesis Test

Example: Should we be drinking Turkish wine?

We all know that France, but what if I told you that Turkish wine is almost as good, at a fraction of the price?!

Some data from Kaggle. Here are top 15 countries in terms of mean point scores:



Two Independent Samples Hypothesis Test

Let's look at some summaries for the top 15:

```
wine %>%
  filter(!is.na(country)) %>%
  group_by(country) %>%
  summarize(Points = mean(points, na.rm
            Price = mean(price, na.rm =
            n = n()) %>%
  top_n(15, Points) %>%
  mutate_at(.vars = vars(Points, Price),
            round, digits = 2) %>%
  arrange(desc(Points)) %>%
  DT::datatable(options = list(pageLengt
                class = "cell-border str
```

England: only 9 observations.

Austria: basically same score as France.

Morocco: only 12 observations.

Turkey: cheaper with good n

	country	Points	Price	n
1	England	92.89	47.5	9
2	Austria	89.28	31.19	3057
3	France	88.93	45.62	21098
4	Germany	88.63	39.01	2452
5	Italy	88.41	37.55	23478
6	Canada	88.24	34.63	196
7	Slovenia	88.23	28.06	94
8	Morocco	88.17	18.83	12
9	Turkey	88.1	25.8	52
10	Portugal	88.06	26.33	5322

Showing 1 to 10 of 15 entries

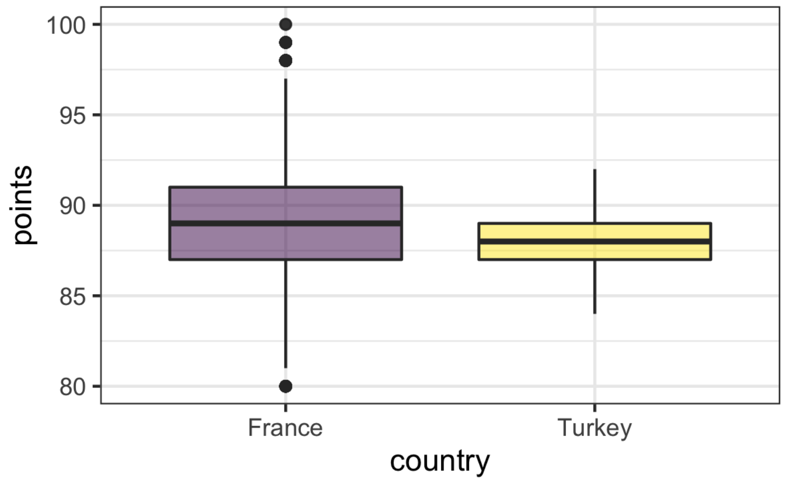
Two Independent Samples Hypothesis Test

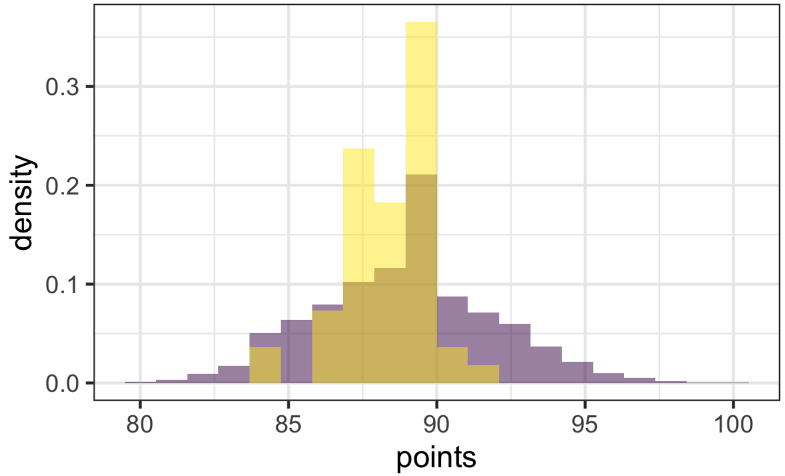
Before you start any sort of analysis, always a good idea to take a look at the data.

```
wine_subset <- wine %>% filter(country %

ggplot(wine_subset,
       aes(x = country, y = points, fill
  scale_fill_viridis_d() + guides(fill =
  geom_boxplot(alpha = 0.5)

ggplot(wine_subset,
       aes(x = points, fill = country))
  scale_fill_viridis_d() + guides(fill =
  geom_histogram(aes(y = after_stat(dens
                 position = "identity",
                 bins = 20)
```





From these plots:

- though means are close, when compared to spreads, could be a difference
- variance do NOT look equal

Two Independent Samples Hypothesis Test

Want to test $H_0 : \mu_{\text{France}} = \mu_{\text{Turkey}}$ against the alternative $H_A : \mu_{\text{France}} \neq \mu_{\text{Turkey}}$. We'll use $\alpha = 0.05$.

We want to do this using a two sample t-test. First, need to check assumptions:

- independent groups
- independent observations
- are averages normally distributed? both n's are greater than 30, so CLT.

Next, need to know if we can assume equal variances. So, find standard deviations for the two groups.

```
wine_subset %>%
  group_by(country) %>%
  summarize(means = mean(points),
            s = sd(points),
            n = n())
```

```
## # A tibble: 2 x 4
##   country means     s     n
## * <chr>   <dbl> <dbl> <int>
## 1 France   88.9  3.20 21098
## 2 Turkey   88.1  1.58    52
```

Since $s_{\text{France}}/s_{\text{Turkey}} > 2$, variances cannot be assumed equal.

Good time to pause and double check with plots: this matches what we saw. Nice!

Two Independent Samples Hypothesis Test

So, we cannot assume equal variances. Hence, we calculate $\widehat{\text{SD}}(V) = \sqrt{\frac{s^2_{\text{France}}}{n_{\text{France}}} + \frac{s^2_{\text{Turkey}}}{n_{\text{Turkey}}}}$, and we then know that **IF** the null hypothesis is true, then $T \sim t_\nu$ where

$$\nu = \frac{\left(\frac{s^2_{\text{France}}}{n_{\text{France}}} + \frac{s^2_{\text{Turkey}}}{n_{\text{Turkey}}}\right)^2}{\frac{(s^2_{\text{France}}/n_{\text{France}})^2}{n_{\text{France}}-1} + \frac{(s^2_{\text{Turkey}}/n_{\text{Turkey}})^2}{n_{\text{Turkey}}-1}}$$

So, let us calculate the two.

```
wine_subset %>%
  group_by(country) %>%
  summarize(s2 = var(points),
            n = n()) %>% print %>%
  summarize(sd_V = sqrt(sum( s2 / n )),
            df = sum( s2 / n )^2 / (sum(
```

```
## # A tibble: 2 x 3
##   country    s2     n
## * <chr>   <dbl> <int>
## 1 France   10.2  21098
## 2 Turkey   2.48     52

## # A tibble: 1 x 2
##    sd_V    df
##   <dbl> <dbl>
## 1 0.220  52.0
```

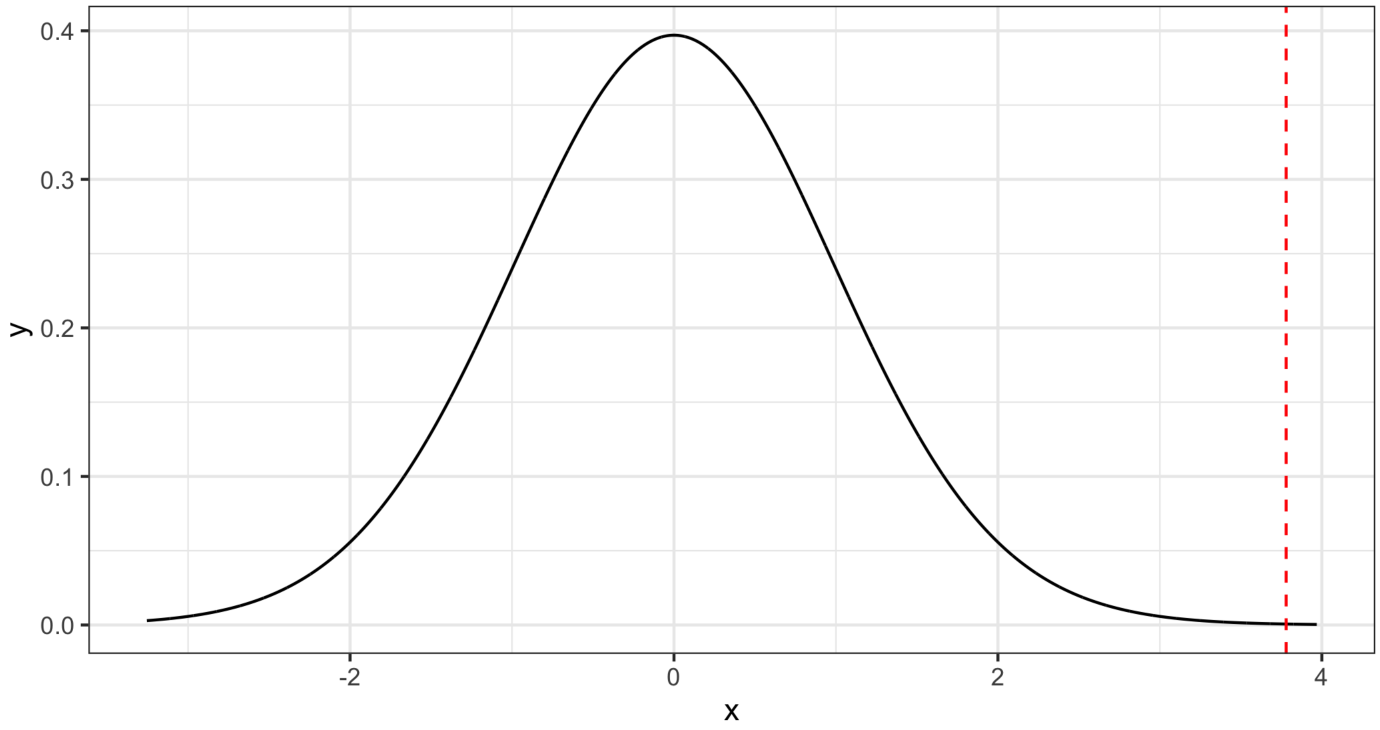
Two Independent Samples Hypothesis Test

So, we can find T_{obs} (I use more digits than previously presented)

```
T_obs <- (88.92587 - 88.09612 - 0)/0.2195276; T_obs
```

```
## [1] 3.779707
```

and compare it to the t-distribution with 52 degrees of freedom:



Two Independent Samples Hypothesis Test

Conclusion using quantiles:

1. find values such that there's $\alpha/2 = 0.025$ to the left and right, respectively:

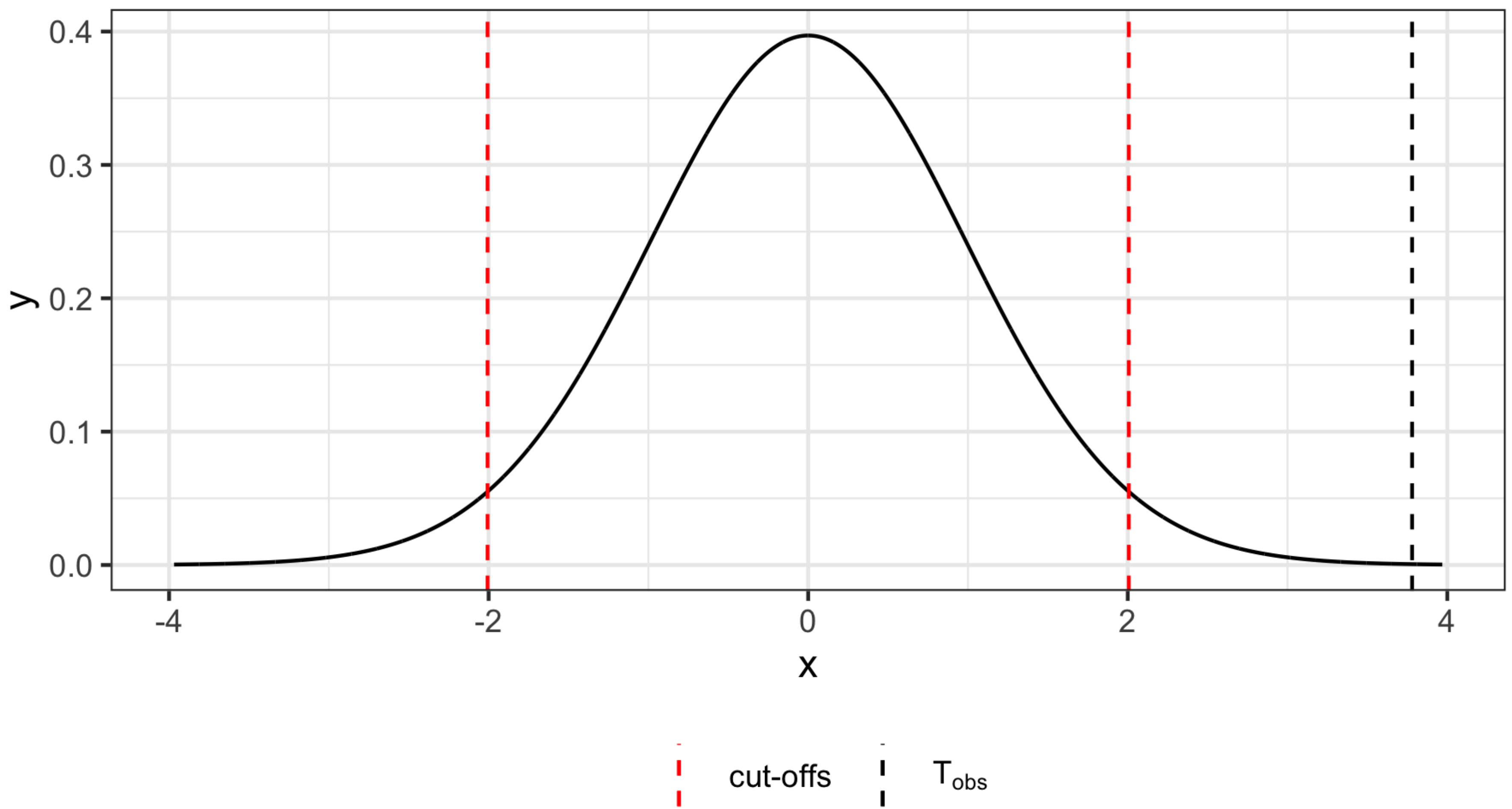
   ```
   T_52 <- StudentsT(df = 52)
   quantile(T_52, c(0.025, 0.975))
   ```

   ```
   ## [1] -2.006647  2.006647
   ```

   these are our cut-offs for when the observed value of the test statistic T_{obs} is far from 0

2. check if our value is outside this interval
   - since we found that $T_{\text{obs}} = 3.779707$, it is outside the cut-offs
3. since T_{obs} is outside the cut-offs, it *is* far from 0
4. since T_{obs} is far from 0, $\bar{X}_{\text{France}} - \bar{X}_{\text{Turkey}}$ *is* far from 0, i.e. $\bar{X}_{\text{France}}$ *is* far from $\bar{X}_{\text{Turkey}}$
5. since $\bar{X}_{\text{France}}$ is far from $\bar{X}_{\text{Turkey}}$, we no longer believe that $\mu_{\text{France}} = \mu_{\text{Turkey}}$.

Two Independent Samples Hypothesis Test



Two Independent Samples Hypothesis Test

Conclusion using the p-value:

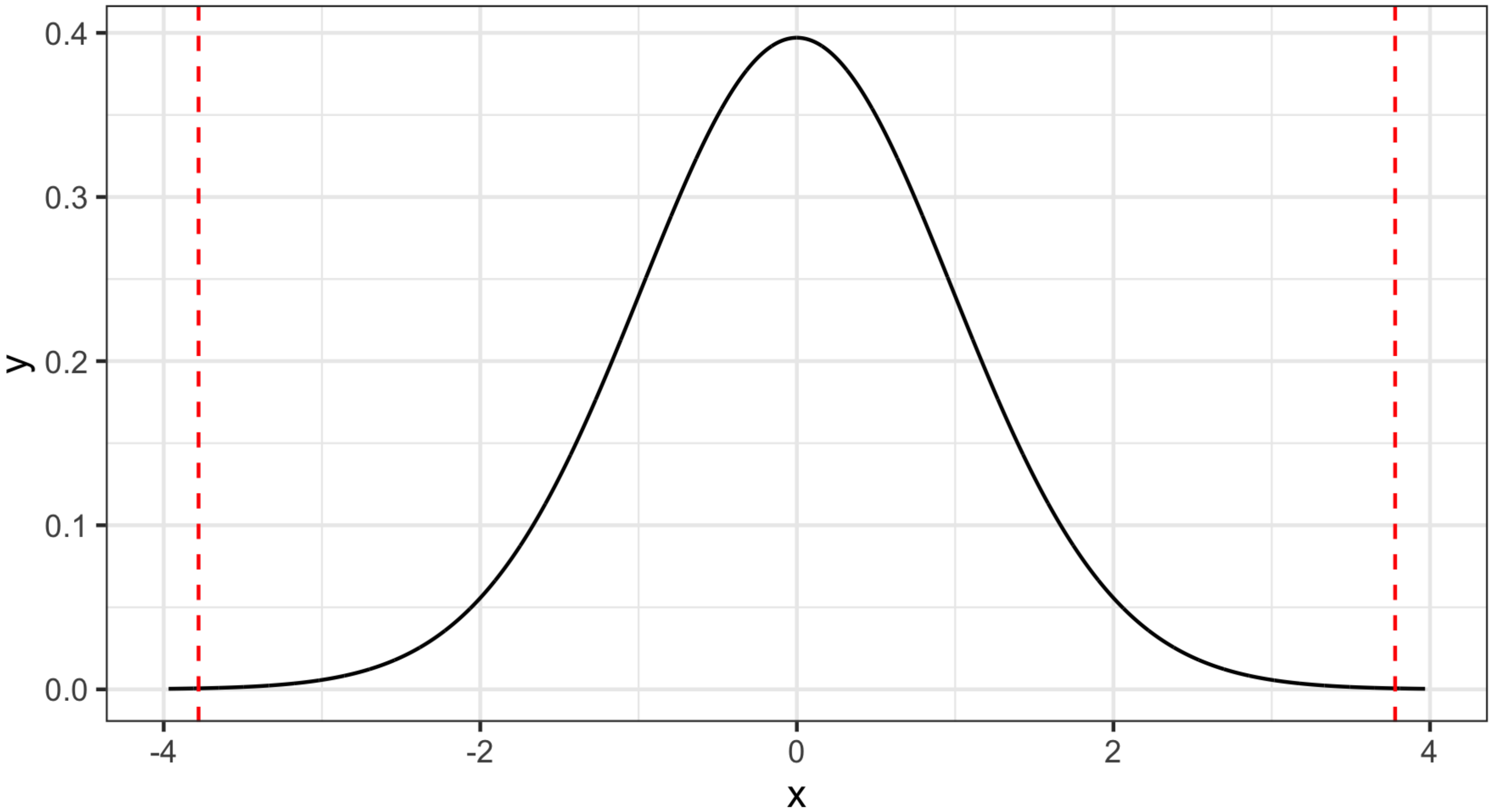
1. find the probability of being further from zero:

```
2*(1 - cdf(T_52, T_obs))
```

```
## [1] 0.0004060735
```

2. since the probability of being further away from zero is less than $\alpha = 0.05$, it *is* small
3. since the probability of being further away from zero is small, T_{obs} *is* far from zero.
4. since T_{obs} is far from 0, $\bar{X}_{\text{France}} - \bar{X}_{\text{Turkey}}$ *is* far from 0, i.e. $\bar{X}_{\text{France}}$ *is* far from $\bar{X}_{\text{Turkey}}$
5. since $\bar{X}_{\text{France}}$ is far from $\bar{X}_{\text{Turkey}}$, we no longer believe that $\mu_{\text{France}} = \mu_{\text{Turkey}}$.

Two Independent Samples Hypothesis Test



Two Independent Samples Hypothesis Test

We can also calculate a 95% confidence interval:

```
wine_subset %>%
  group_by(country) %>%
  summarize(means = mean(points),
            s = sd(points),
            n = n())
```

```
## # A tibble: 2 x 4
##   country means     s     n
## * <chr>   <dbl> <dbl> <int>
## 1 France   88.9  3.20 21098
## 2 Turkey   88.1  1.58    52
```

$$\hat{V} \pm t_{52,0.05/2}\widehat{\mathrm{SD}}(V) = (88.92587 - 88.09615) \pm 2.006647\sqrt{\frac{3.199695^2}{21098} + \frac{1.575046^2}{52}}$$

$$= 0.82972 \pm 0.4405144$$

So, we are 95% confident that the true difference in mean points for wines from France vs wines from Turkey is in the interval [0.39, 1.27].

Two Independent Samples Hypothesis Test

Using `t.test` to double check our results:

```
t.test(data = wine_subset,
       points ~ country, var.equal = FALSE)
```

```
##
##      Welch Two Sample t-test
##
## data:  points by country
## t = 3.7796, df = 52.043, p-value = 0.000406
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3892102 1.2702216
## sample estimates:
## mean in group France mean in group Turkey
##             88.92587             88.09615
```

Two Independent Samples Hypothesis Test

Revisit: Corona Virus data

Previously, we considered the death rates in Italy and China. Back then, we didn't have the tools to actually test the hypothesis that they are the same.

Let's consider the hypothesis $H_0 : \pi_{\text{Italy}} = \pi_{\text{China}}$, and test it against $H_A : \pi_{\text{Italy}} \neq \pi_{\text{China}}$ using $\alpha = 0.01$.

Most recent data:

```
corona_subset
```

```
## # A tibble: 2 x 5
##   Country deaths confirmed recovered  p_hat
## * <chr>    <dbl>     <dbl>     <dbl>  <dbl>
## 1 China     3240     79243     71060 0.0409
## 2 Italy     6820     69176      8326 0.0986
```

Two Independent Samples Hypothesis Test

We will reject null if P_{Italy} is far from P_{China}. So, we will consider $P_{\text{Italy}} - P_{\text{China}}$.

To be able to say if they are "far from each other", we want to find the probability that they are further from each other. So, we need the distribution of some quantity we know.

Remember, we also do this assuming H_0 is true.

Remember, P_{Italy} and P_{China} are proportions, so approximately normally distributed if the n's are large enough.

So, if the n's are large enough, the difference will be normally distributed with mean 0. But what is the variance?

If H_0 is true, then $\pi_{\text{Italy}} = \pi_{\text{China}}$. Let's call this common proportion π_0. So,

$$\mathrm{Var}(P_{\text{Italy}}) = \frac{\pi_{\text{Italy}}(1-\pi_{\text{Italy}})}{n_{\text{Italy}}} = \frac{\pi_0(1-\pi_0)}{n_{\text{Italy}}}$$
$$\mathrm{Var}(P_{\text{China}}) = \frac{\pi_{\text{China}}(1-\pi_{\text{China}})}{n_{\text{China}}} = \frac{\pi_0(1-\pi_0)}{n_{\text{China}}}.$$

Two Independent Samples Hypothesis Test

Therefore, assuming the two groups are independent,

$$\mathrm{Var}(P_{\text{Italy}} - P_{\text{China}}) = \mathrm{Var}(P_{\text{Italy}}) + \mathrm{Var}(P_{\text{China}}) = \pi_0(1-\pi_0)\left(\frac{1}{n_{\text{Italy}}} + \frac{1}{n_{\text{China}}}\right).$$

So, if the null hypothesis is true,

$$P_{\text{Italy}} - P_{\text{China}} \sim N\left(0, \pi_0(1-\pi_0)\left(\frac{1}{n_{\text{Italy}}} + \frac{1}{n_{\text{China}}}\right)\right),$$

or equivalently,

$$Z = \frac{P_{\text{Italy}} - P_{\text{China}}}{\sqrt{\pi_0(1-\pi_0)\left(\frac{1}{n_{\text{Italy}}} + \frac{1}{n_{\text{China}}}\right)}} \sim N(0,1).$$

Notice how this is of the form $\frac{V - v_0}{\widehat{\mathrm{SD}}(V)}$ **IF** the null hypothesis is true.

Two Independent Samples Hypothesis Test

There are a few questions we need to answer:

1. how big do $n_{\text{Italy}}, n_{\text{China}}$ have to be?
2. how should we estimate π_0?

To answer the first question: we need the sample sizes big enough that P_{Italy} and P_{China} are approximately normally distributed when the null hypothesis is true. Previously, we said that if $\pi_{\text{Italy}} n_{\text{Italy}} > 5$ and $(1 - \pi_{\text{Italy}}) n_{\text{Italy}} > 5$, then all is well. (Same for n_{China}.)

Will use same rule of thumb here. But remember, when H_0 is true, $\pi_{\text{Italy}} = \pi_{\text{China}} = \pi_0$. So, we need

$$\pi_0 n_{\text{Italy}} > 5 \quad \text{and } (1 - \pi_0) n_{\text{Italy}} > 5$$

and

$$\pi_0 n_{\text{China}} > 5 \quad \text{and } (1 - \pi_0) n_{\text{China}} > 5$$

Two Independent Samples Hypothesis Test

What should we use to estimate π_0? Well, if the null hypothesis is true, the two groups are basically the same (in terms of the true proportion, at least). So, to estimate π_0, we will treat the two groups as one big group:

$$P_0 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{n_{\text{Italy}}P_{\text{Italy}} + n_{\text{China}}P_{\text{China}}}{n_{\text{Italy}} + n_{\text{China}}}$$

Two Independent Samples Hypothesis Test

So, first we will find our observed value of P_0:

```
corona_subset %>% summarize(p_0 = sum(p_hat*confirmed)/(sum(confirmed)))
```

```
## # A tibble: 1 x 1
##      p_0
##    <dbl>
## 1 0.0678
```

Next, we check if the two sample sizes are big enough:

```
corona_subset %>%
  mutate(p_0 = sum(p_hat*confirmed)/(sum(confirmed)),
         check_n1 = confirmed*p_0,
         check_n2 = confirmed*(1-p_0))
```

```
## # A tibble: 2 x 8
##   Country deaths confirmed recovered  p_hat    p_0 check_n1 check_n2
## * <chr>    <dbl>     <dbl>     <dbl>  <dbl>  <dbl>    <dbl>    <dbl>
## 1 China     3240     79243     71060 0.0409 0.0678    5371.   73872.
## 2 Italy     6820     69176      8326 0.0986 0.0678    4689.   64487.
```

Two Independent Samples Hypothesis Test

Finally, calculate the observed value of our test statistic:

$$Z_{\text{obs}} = \frac{0.09858911 - 0.04088689}{\sqrt{0.06778108 \cdot (1 - 0.06778108)\left(\frac{1}{69176} + \frac{1}{79243}\right)}} = 44.1156569$$

We compare this to the standard normal.

Two Independent Samples Hypothesis Test

Conclusion using quantiles: is our observed value "far from 0"? Find cut-offs such that only $\alpha/2 = 0.005$ is further from 0 on each side:

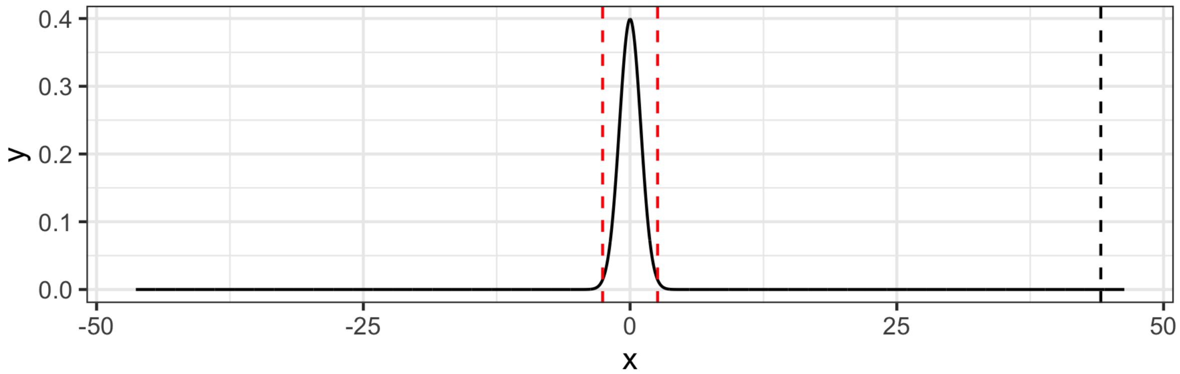
```
quantile(Normal(), c(0.005, 0.995))
```

```
## [1] -2.575829  2.575829
```

Since we observed 44.1156569, we observe something far from 0.

So, P_{Italy} is far from P_{China}.

So, we reject the idea that $\pi_{\text{Italy}} = \pi_{\text{China}}$.



Two Independent Samples Hypothesis Test

As always, we are probably more interested in a confidence interval than a hypothesis test. Fortunately, we can "easily" create that. However, we need to take a step back.

We would like to simply take our test statistic, $\frac{P_{\text{Italy}} - P_{\text{China}}}{\sqrt{\pi_0(1-\pi_0)\left(\frac{1}{n_{\text{Italy}}} + \frac{1}{n_{\text{China}}}\right)}}$, and rearrange it. However, this only follows a normal distribution **IF** the null hypothesis is true.

Therefore, we need to be a bit more general, and instead use $\frac{P_{\text{Italy}} - P_{\text{China}}}{\sqrt{\left(\frac{P_{\text{Italy}}(1-P_{\text{Italy}})}{n_{\text{Italy}}} + \frac{P_{\text{China}}(1-P_{\text{China}})}{n_{\text{China}}}\right)}}$. We can then construct a $(1-\alpha)\%$ CI as

$$P_{\text{Italy}} - P_{\text{China}} \pm z_{\alpha/2}\sqrt{\left(\frac{P_{\text{Italy}}(1-P_{\text{Italy}})}{n_{\text{Italy}}} + \frac{P_{\text{China}}(1-P_{\text{China}})}{n_{\text{China}}}\right)}.$$

Notice, this is of the form $V \pm z_{\alpha/2}\widehat{\text{SD}}(V)$.

We use Z instead of T here, since $V \sim N$, and we only do not have to estimate the standard deviation separate from estimating the means.

Two Independent Samples Hypothesis Test

```r
corona_subset %>% print %>%
  summarize(LL = p_hat[2] - p_hat[1] - 2.576*sqrt(sum(p_hat*(1-p_hat)/confirmed)),
            UL = p_hat[2] - p_hat[1] + 2.576*sqrt(sum(p_hat*(1-p_hat)/confirmed)))
```

```
## # A tibble: 2 x 5
##   Country deaths confirmed recovered  p_hat
## * <chr>    <dbl>     <dbl>     <dbl>  <dbl>
## 1 China     3240     79243     71060 0.0409
## 2 Italy     6820     69176      8326 0.0986

## # A tibble: 1 x 2
##       LL     UL
##    <dbl>  <dbl>
## 1 0.0543 0.0611
```

Two Independent Samples Hypothesis Test

We can do this in R using the function prop.test:

```
prop.test(x = c(6820, 3240), n = c(69176, 79243), correct = FALSE, conf.level = 0.99)
```

```
##
##     2-sample test for equality of proportions without continuity
##     correction
##
## data:  c(6820, 3240) out of c(69176, 79243)
## X-squared = 1946.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.05426606 0.06113837
## sample estimates:
##     prop 1     prop 2
## 0.09858911 0.04088689
```

Important things to know about prop.test:

- it uses Z^2 as the test statistic instead of Z
  - notice how $\sqrt{1946.2} = 44.1157568 \approx 44.1156569$
- you need to specify `correct = FALSE`