Math 300 Lesson 40 Notes

Inference Examples

YOUR NAME HERE

July, 2022

Contents

Objectives

1. Analyze and interpret hypothesis tests and confidence intervals using traditional methods and non-traditional methods.

Reading

Appendix B

Lesson

- Complete an analysis for a one sample test of means. Compare and contrast traditional and non-traditional methods. Be able to due this for other tests.
- The map provides a visual representation of the ideas and gives examples of problems and code. It is worth spending some time on this map.
- Review for the final. Many steps in this lesson do help prepare for the final. We are introducing some new ideas that are not going to be on the final but are relevant to know.

Libraries

```
library(tidyverse)
library(infer)
library(moderndive)
```

Problem Statement

The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. 5,534 randomly sampled US women between 2006 and 2010 completed the survey. The women sampled here had been married at least once. Do we have evidence that the mean age of first marriage for all US women from 2006 to 2010 is greater than 23 years?

Hypothesis Statement

In words:

Null hypothesis: The mean age of first marriage for all US women from 2006 to 2010 is equal to 23 years.

Alternative hypothesis: The mean age of first marriage for all US women from 2006 to 2010 is greater than 23 years.

In symbols (with annotations)

$$H_O : \mu = \mu_0$$

where $/mu$ represents the mean age of first marriage for all US women from 2006 to 2010 and $/mu_0$ is 23.

$$H_A : \mu > 23$$

This is a one-sided hypothesis test. We will set α equal to 0.05.

Exploratory

Let's get the data and explore them.

```
age_at_marriage <- read_csv("https://moderndive.com/data/ageAtMar.csv")
```

```
head(age_at_marriage)
```

```
## # A tibble: 6 x 1
##     age
##   <dbl>
## 1    32
## 2    25
## 3    24
## 4    26
## 5    32
## 6    29
```

```
# Complete the code
# age_at_marriage %>%
#   summarize(
#     sample_size = ____________,
#     mean = _______________,
#     sd = _____________,
#     minimum = _________________,
#     lower_quartile = ____________________,
#     median = ___________________,
#     upper_quartile = _________________________,
#     max = max(age)
#   )
```

```
# Complete the code
# ggplot(data = age_at_marriage, mapping = aes(x = _____________)) +
#   geom_histogram(binwidth = 3, color = "black", fill = "cyan") +
#   theme_classic()
```

Does the data appear to be normally distributed?

Non-traditional methods

Since we only have a single variable, it is hard to imagine how we could do a permutation test since we don't have a second variable as a label to shuffle. One idea is to make an additional assumption of symmetry, which may not be appropriate for this problem. If we make that assumption, under the null half the values should be greater than 23 and half should be less. We could create a new variable by subtracting the hypothesized mean. We then count how many are positive as a test statistic. We then randomly select positive and negative values to find the sampling distribution. This may be confusing, so let's write the code out. If the assumption of symmetry is not valid, then the test is a test about the median and not the mean.

We have to worry about those values that are exactly 23. Let's find out how big of a problem that is going to be.

```
age_at_marriage %>%
  mutate(delta=age-23,diff_sign=if_else(delta<0,-1,1)) %>%
  pull(delta) %>%
  table()
```

```
## .
## -13 -11 -10  -9  -8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4   5   6   7
##   1   2   5  23  52 106 155 368 468 474 495 480 452 419 427 334 262 227 203 144
##   8   9  10  11  12  13  14  15  16  17  18  19  20
## 104  69  74  52  43  18  23  15  13  11   7   7   1
```

There are 452 observations that are exactly 23. We can randomly make half positive and half negative. They cancel each other out so we don't need to count them in our code.

```
age_at_marriage %>%
  filter(age!=23) %>%
  mutate(delta=age-23,n=n(),diff_sign=if_else(delta<0,-1,1)) %>%
  summarize(test_stat=sum(diff_sign))
```

```
## # A tibble: 1 x 1
##   test_stat
##       <dbl>
## 1      -176
```

We have 176 more ages that were less than 23 than we had greater than 23. If the assumption of symmetry is appropriate, we can sample we replacement from a vector of -1 and 1 and count the total. This would give us the null distribution. Mathematically we could make this easier using the binomial distribution but that is beyond the scope of this class. Note that under the alternative hypothesis, we can't reject with this test since we have more less than 23 than we do greater than 23.

```
# Complete the code
set.seed(7620)
# Repeat resampling 10000 times
# virtual_resamples <- tibble(value = c(-1,1)) %>%
#   rep_sample_n(size = _________, replace = TRUE, reps = ____________)
```

```
# Complete the code
# Compute 10000 sample statistics
# samp_dist <- virtual_resamples %>%
#   group_by(_____________) %>%
#   summarize(stat_count = sum(_______________))
```

```
# Complete the code
# samp_dist %>%
#   ggplot(aes(x=_________________)) +
#   geom_histogram(bins=70,color="black",fill="cyan") +
#   theme_classic()
```

The p-value is the probability of our observed value or more extreme, which for this one-sided test means greater than.

```
# Complete the code
# samp_dist %>%
#   summarize(p_value=sum(______________>=-176)/10000)
```

Conclusion:

Let's do this differently using the bootstrap.

From the book:

In order to look to see if the observed sample mean of 23.44 is statistically greater than 23, we need to account for the sample size. We also need to determine a process that replicates how the original sample of size 5534 was selected.

We can use the idea of bootstrapping to simulate the population from which the sample came and then generate samples from that simulated population to account for sampling variability. Recall how bootstrapping would apply in this context:

- Sample with replacement from our original sample of 5534 women and repeat this process 10,000 times

- calculate the mean for each of the 10,000 bootstrap samples created in Step 1.,

- combine all of these bootstrap statistics calculated in Step 2 into a `boot_distn` object, and
- shift the center of this distribution over to the null value of 23. (This is needed since it will be centered at 23.44 via the process of bootstrapping.)

```
# Complete the code
set.seed(2018)
# null_distn_one_mean <- _______________ %>%
#   specify(response = ____________) %>%
#   hypothesize(null = "point", mu = _____________) %>%
#   generate(reps = 10000, type="____________________") %>%
#   calculate(stat = "___________")
```

```
# Complete the code
# ___________________ %>% visualize()
```

Now the p-value.

```
# Complete the code
# null_distn_one_mean %>%
#   visualize() +
#   shade_p_value(obs_stat = 23.44019   , direction = "_____________________")
```

```
# Complete the code
# null_distn_one_mean %>%
#   get_pvalue(obs_stat = 23.44019, direction = "_____________________")
```

Conclusion:

Confidence interval

```
# Complete the code
# boot_distn_one_mean <- ________________ %>%
#   specify(response = _____________) %>%
#   generate(reps = 10000, type="__________________") %>%
#   calculate(stat = "_________________")
```

```
# Complete the code
# boot_distn_one_mean %>%
#   get_ci(level=_______________)
```

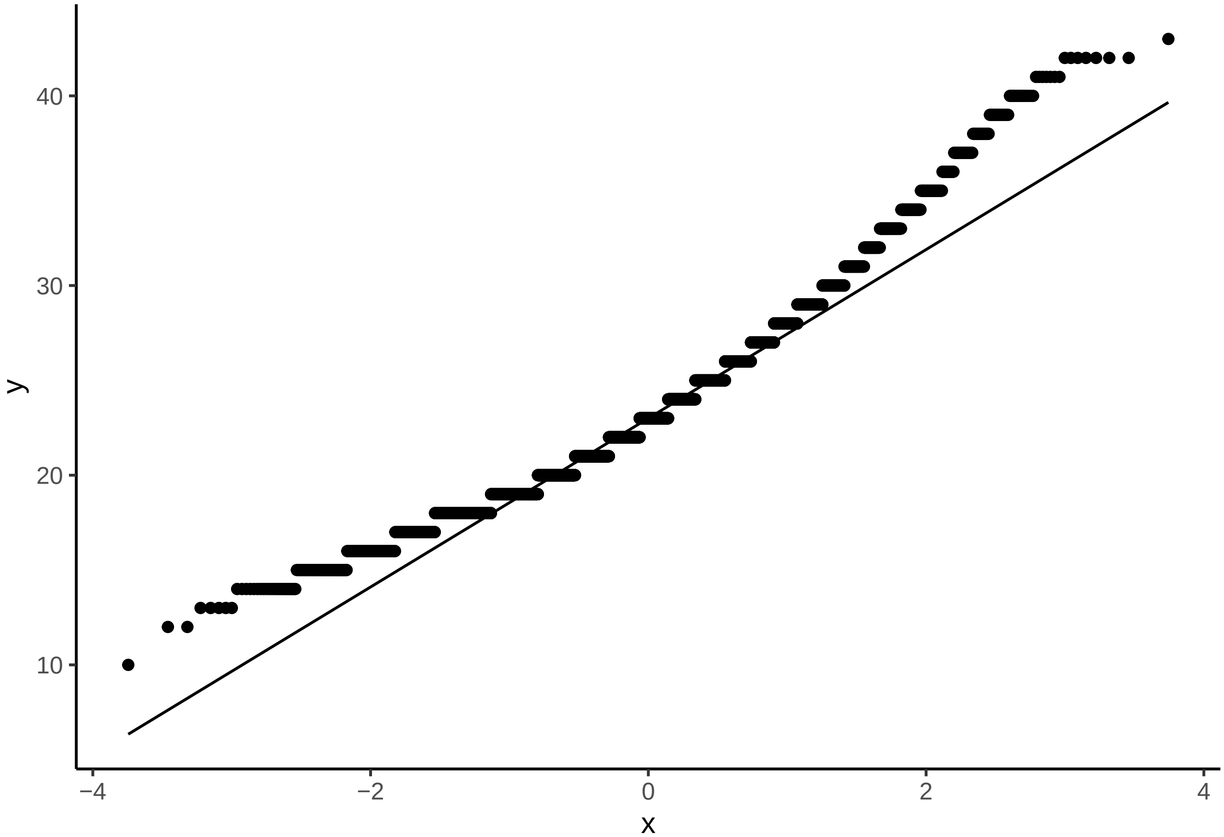
Conclusion:

Traditional Methods

To use a t-test we require the data to be normally distributed and independent. The independence is difficult to check and depends on the sampling. Since the data was collected from a random sample, independence seems reasonable.

The normality assumption is also difficult to check. The histogram indicates that the data has some skewness. We can also use a *qq* plot.

```
ggplot(data = age_at_marriage, mapping = aes(sample = age)) +
  stat_qq() +
  stat_qq_line() +
  theme_classic()
```



From this plot we see that small values are not as small as they should be if normality applied. The points are above the line. For the large values they are larger than they should be under normality as the points are above the line. This indicates a right skewness.

Since the data set is so large, the lack of symmetry is less of a problem so we will continue with the traditional method.

```
t_test(x=age_at_marriage,formula=age~NULL,mu=23,alternative="greater")
```

```
## # A tibble: 1 x 7
##   statistic  t_df  p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      6.94  5533 2.25e-12 greater         23.4     23.3      Inf
```

We, therefore, have sufficient evidence to reject the null hypothesis. Based on this sample, we have evidence that the mean age of first marriage for all US women from 2006 to 2010 is greater than 23 years.

The bootstrap and traditional method agree. The randomization test did not support because the lack of symmetry was an issue.

Documenting software

- File creation date: 2022-07-06
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4
- `infer` package version: 1.0.0