Topic 2.3: Basics of statistical inference - hypothesis tests

"Introduction to testing statistical significance: weighing evidence from data"

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Context

As we learned in the first half of the semester, *statistics* are functions of random variables and therefore are random variables themselves. In particular, they have their own distributions, called *sampling distributions*. Our inferred knowledge about these distributions is used to *estimater parameters* of the model which we postulate was used to generate the data.

So far we have learned about *point estimators*. Next, we will learn about *interval estimators*. And now, we focus on hypothesis tests. In other notes, we discuss the general view of how these fit together.

You may see in statistics courses you take that hypothesis tests are very useful in the addressing the question *whether the postulated model was indeed used to generate the data*.

Goals of this lecture

- Review of what we have learned in statistical inference
  - Estimation basics (statistics as simple point estimators)
  - Confidence intervals
  - Hypothesis tests
  - p-values
- Where have we seen p-values before? [Let's come full circle!]
- Let's see a *full story* example, from scratch, for a discrete distribution: it provides a full overview, as well as a basis for some final exam questions.

Reminder: interpretation of confidence intervals

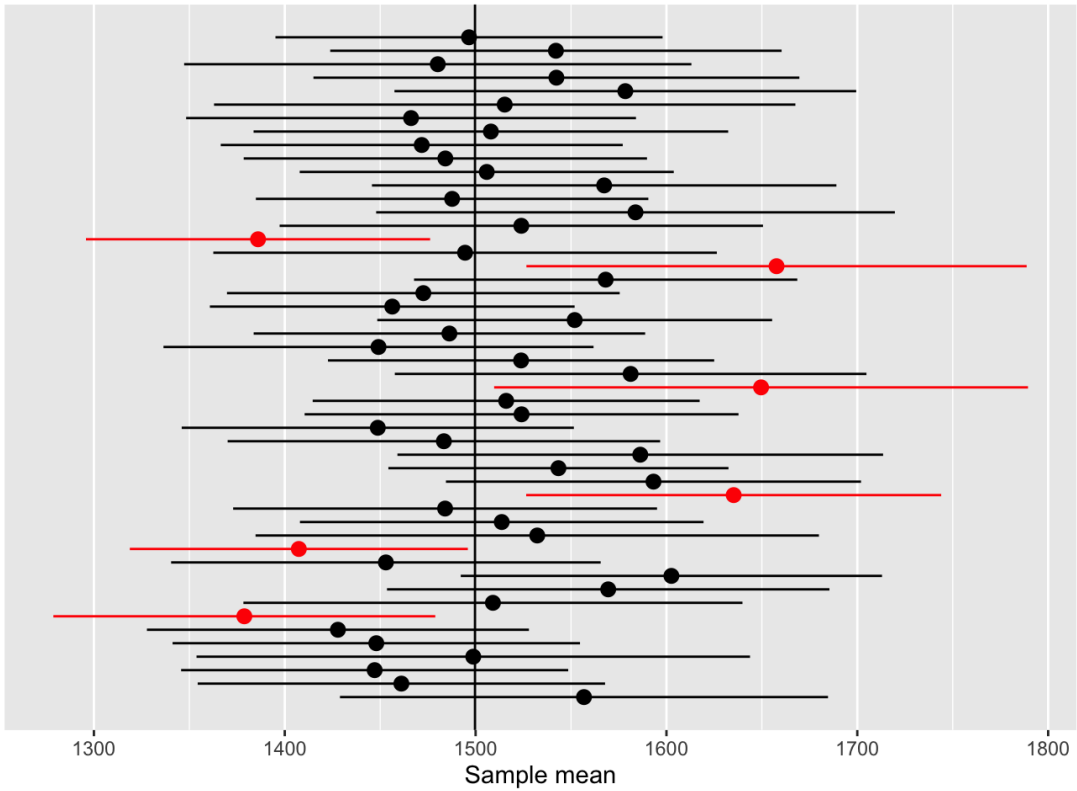


Figure 1: 100 confidence intervals for Ames housing

These are more informative than point estimators, yes.

What is the weight of the evidence provided by your data?

"I'm not confortable just giving a confidence interval.."

- What if we could 'measure' how 'significant' the variability observed is?

Enter live lecture.

Outside of these slides, we now go over the following topics:

- basic elements of a statistical test
- p-value. . . . wait, does this look familiar?
  - Revisit previous case studies from a new point of view!
- relationship between confidence intervals and a hypothesis test?
- how do you know if the test is one-sided or two-sided?

Overview of hypothesis testing

Elements of a hypothesis test

- hypotheses (working(null) and research(alternative))
- a test statistic (that can measure disrepancy related to the hypotheses)
- a rejection region (used to create the formal decision rule)

Of course there is more...

- Errors
  - (what? remember, we are never certain about anything when dealing with data... we merely have probabilities, and we are trying to quantify the uncertainty formally!)
- Advanced topics:
  - power of a test,
  - 'best' tests,
  - testing model goodness of fit,
  - my data isn't usual! help! there's no model!,
  - etc.

Revisiting some old friends!

Recall Case study 1: automobile parts

Problem.

An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produce parts having a mean diameter of 5.0 millimeters. The engineer involved claims that the population mean is 5.0 millimeters.

An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation is $\sigma = 0.1$ millimeter. The experiment indicates a sample average diameter of $\overline{X} = 5.027$ millimeters.

Question:

Does this sample information appear to support or refute the engineer's claim?

Recall Case study 1: Solution

$$P(|\overline{X} - 5| \geq 0.027) = 2P\left(\frac{\overline{X} - 5}{0.1\sqrt{100}} \geq 2.7\right) = 0.0035 = 0.007$$

Anything familiar?

Recall Case study 1: Solution

$$P(|\overline{X} - 5| \geq 0.027) = 2P\left(\frac{\overline{X} - 5}{0.1\sqrt{100}} \geq 2.7\right) = 0.0035 = 0.007$$

Anything familiar?

- What is this probability?
  - Prob of seeing the observed data or more extreme. . . ..
  - under some assumption about the population mean!
- So 0.007 is a p-value!

$$H_0 : \mu = 5.0 \text{ vs. } H_1 : \mu \neq 5.0$$

Recall Case Study 2: paint drying time

Problem Two independent experiments are run in which two different types of paint are compared. 18 specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Question:

Assuming that the mean drying time is equal for the two types of paint, find $P(\overline{X}_A - \overline{X}_B > 1)$, where $\overline{X}_A$ and $\overline{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

Recall Case study 2: Solution - solution

The probability that we compute is given by:

$$P(\overline{X}_A - \overline{X}_B > 1) = P\left(\frac{\overline{X}_A - \overline{X}_B - 0}{\sqrt{1/9}} \geq \frac{1-0}{\sqrt{1/9}}\right) = P(Z > 3) = 0.0013.$$

Anything familiar?

Recall Case study 2: Solution - solution

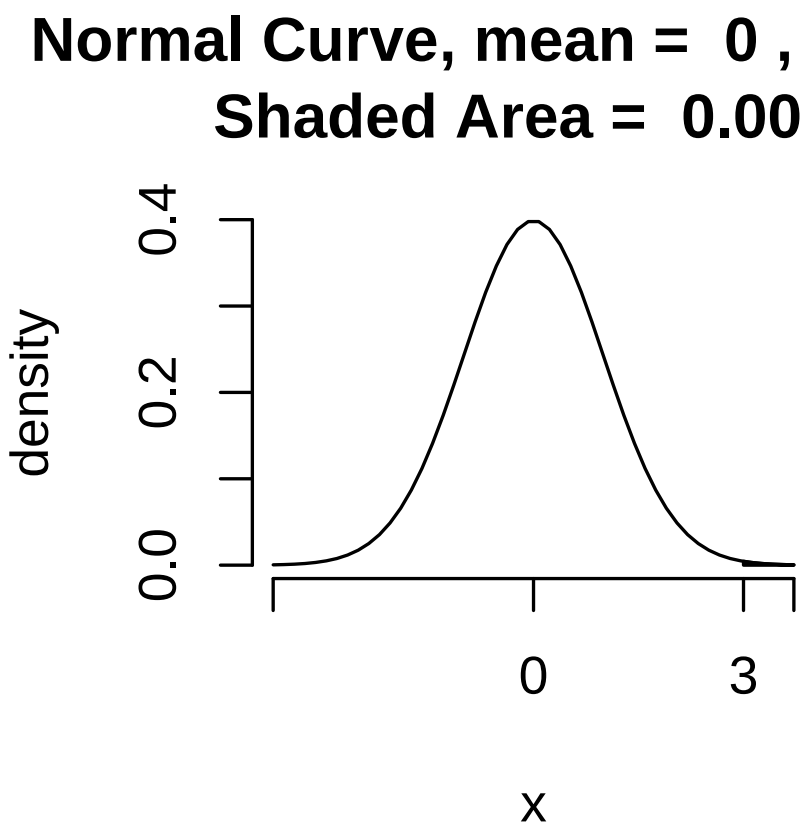
The probability that we compute is given by:

$$P(\overline{X}_A - \overline{X}_B > 1) = P\left(\frac{\overline{X}_A - \overline{X}_B - 0}{\sqrt{1/9}} \geq \frac{1-0}{\sqrt{1/9}}\right) = P(Z > 3) = 0.0013.$$

Anything familiar?

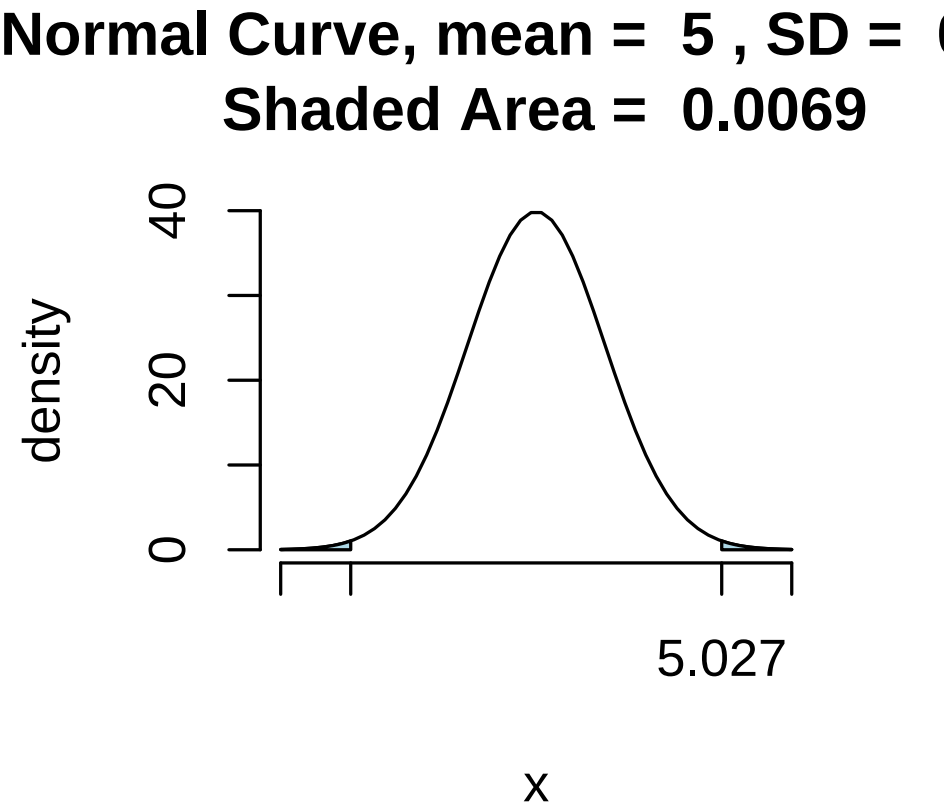
- What is this probability?
  - Prob of seeing the observed data or more extreme. . . .
  - under some assumption about two population means!
- So 0.0013 is a p-value again!
- $$H_0 : \mu_A - \mu_B = 0 \text{ vs. } H_1 : \mu_A - \mu_B > 0.$$

```
## Case study 1: using the z-value
pnormGC(3, region="above", mean=0, sd=1,graph=TRUE)
```




```
[1] 0.001349898
```

```
## Case study 2
pnormGC(bound=c(4.973,5.027), region="outside",
        mean=5, sd=0.1/sqrt(100),graph=TRUE)
```




```
[1] 0.006933948
```

Discrete data: tests for population proportion

Example 1: estimating a proportion

In a random sample of n = 500 families owning television sets in the city of Hamilton, Canada, it is found that x = 340 subscribe to HBO. Find a 95% confidence interval for the actual proportion of families with television sets in this city that subscribe to HBO.

- random variable: number of...
- unknown population parameter: p, proportion
- statistic (estimate): $\hat{p}$, sample proportion

  What is the sampling distribution of this statistics? What is the setup here?

(If you know this, then you know what `R` function to call to compute the confidence interval!)

Here is a general (approximate) test for equal or given proportions:

```
prop.test(350,500)
```

```
    1-sample proportions test with continuity correction

data:  350 out of 500
X-squared = 79.202, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.6574021 0.7394725
sample estimates:
  p
0.7
```

```r
help(prop.test,package="stats")
prop.test(350,500,p=0.5,alternative="two.sided",conf.level = .
```

```
    1-sample proportions test with continuity correction

data:  350 out of 500
X-squared = 79.202, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
98 percent confidence interval:
 0.6493563 0.7462713
sample estimates:
  p
0.7
```

```r
# alternative can be "two.sided" or "greater" or "less"
```

Some simple examples from R's help file

Compare the following things:

```
heads <- rbinom(1, size = 100, prob = .5)
prop.test(heads, 100)          # continuity correction TRUE by default
```

```
    1-sample proportions test without continuity correction

data:  heads out of 100
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4038315 0.5961685
sample estimates:
  p
0.5
```

Some simple examples from R's help file

Exact test:

```
binom.test(heads,100) # an exact test of a binomial hypothesis!
```

```
data:  heads out of 100
number of successes = 50, number of trials = 100, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3983211 0.6016789
sample estimates:
probability of success
                   0.5
```

Theorems!

- Goal: the sampling distribution of $\hat{p}$
- Assumptions: the unknown proportion p is not expected to be too close to 0 or 1
- Method:
  - Designating a failure in each binomial trial by the value 0 and a success by the value 1,
  - the number of successes, x, can be interpreted as the sum of n values consisting only of 0 and 1s.
  - Then, $\hat{p}$ is just the sample mean of these n values.
  - Hence, by the **Central Limit Theorem**, for n sufficiently large, $\hat{p}$ is **approximately normally distributed**!

What is the z-statistic?

(... figure out the missing pieces.)

Live lecture notes..

Now we discuss *large-sample* intervals and tests for p.

- Some difficulties setting up the z-test by hand.
- Hence the use of R's built-in functions `binom.test()` and `prop.test()`.
- Exact vs. approximate tests: usually we rely on approximations, as long as we know they are fairly accurate.
  - I resort to exact tests when I face a scenario where the assumptions underlying approximation results are questionable. (More on that topic in some other statistics courses.)

Example 2: hypothesis test for a proportion

A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief.

Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? (NO/YES: Use a 0.05 level of significance.)

Example 2: hypothesis test for a proportion

A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief.

Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? (NO/YES: Use a 0.05 level of significance.)

```
prop.test(x=70,n=100,p=0.6,alternative="greater")
```

```
    1-sample proportions test with continuity correction

data:  70 out of 100
X-squared = 3.7604, df = 1, p-value = 0.02624
alternative hypothesis: true p is greater than 0.6
95 percent confidence interval:
 0.6149607 1.0000000
sample estimates:
  p
0.7
```

What are tests of goodness of fit?

What if...

... we are looking at a relationship between two categorical variables?

	mother.smokes	
birthwt.below.2500	no	yes
no	86	44
yes	29	30

Contingency table.

The *research hypothesis* can be made about a model: for example, since it looks like there's a positive association between low birthweight and smoking status, maybe the data is enough evidence to refute H_0: *weight and smoking are independent*!

Testing model fit?

To test for significance, we just need to pass our 2×2 table into the appropriate function. Here's the result of using Fisher's exact test by calling `fisher.test`

```
birthwt.fisher.test <- fisher.test(weight.smoke.tbl)
birthwt.fisher.test
```

```
    Fisher's Exact Test for Count Data

data:  weight.smoke.tbl
p-value = 0.03618
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.028780 3.964904
sample estimates:
odds ratio
  2.014137
```

Testing model fit another way

You can also use the chi-squared test via the `chisq.test` function.

```
chisq.test(weight.smoke.tbl)
```

```
    Pearson's Chi-squared test with Yates' continuity correcti

data:  weight.smoke.tbl
X-squared = 4.2359, df = 1, p-value = 0.03958
```

You get essentially the same answer by running the chi-squared test, but the output isn't as useful. In particular, you're not getting an estimate or confidence interval for the odds ratio. This is why I prefer fisher.test() for testing 2 x 2 tables.

Tests for $j \times k$ tables

Here's a small data set on party affiliation broken down by gender.

```
# Manually enter the data
politics <- as.table(rbind(c(762, 327, 468),
                           c(484, 239, 477)))
dimnames(politics) <- list(
  gender = c("F", "M"),
  party = c("Democrat","Independent", "Republican"))

politics # display the data
```

```
      party
gender Democrat Independent Republican
     F      762         327        468
     M      484         239        477
```

We may be interested in asking whether men and women have different party affiliations.

The answer will be easier to guess at if we convert the rows to show proportions instead of counts. Here's one way of doing this.

```
politics.prop <- prop.table(politics, 1)
politics.prop
```

```
      party
gender  Democrat Independent Republican
     F 0.4894027   0.2100193  0.3005780
     M 0.4033333   0.1991667  0.3975000
```

By looking at the table we see that Female are more likely to be Democrats and less likely to be Republicans.

We still want to know if this difference is significant. To assess this we can use the chi-squared test (on the counts table, not the proportions table!).

```
chisq.test(politics)
```

```
    Pearson's Chi-squared test

data:  politics
X-squared = 30.07, df = 2, p-value = 2.954e-07
```

There isn't really a good one-number summary for general j x k tables the way there is for 2 x 2 tables. One may collapse or use other strategies.

End of course module

Summary of hypothesis testing

- Three elements of a test: hypotheses, test statistic, and rejection region/decision rule
- In practice, check assumptions to know which test to use (i.e., which distribution to reference)

Summary of hypothesis testing

- Three elements of a test: hypotheses, test statistic, and rejection region/decision rule
- In practice, check assumptions to know which test to use (i.e., which distribution to reference)
- We learned about: one- and two-population location and scale problems, in continuious setting, and proportion in discrete setting.

Summary of hypothesis testing

- Three elements of a test: hypotheses, test statistic, and rejection region/decision rule
- In practice, check assumptions to know which test to use (i.e., which distribution to reference)
- We learned about: one- and two-population location and scale problems, in continuious setting, and proportion in discrete setting.
- Hypothesis tests can be made for the general setup: *"Is the model correct? does it fit my data?"* which is a question you should ask before using the model for analytics.
  - Sometimes model fit tests are done heuristically;
  - Sometimes we have formal testing procedures.
  - So far we've only seen a simple two-way table example of this. More to come within regression.

Pro-tips

Let's take a look at a simulated example, for those who want to understand the testing setup a bit more.

What is statistical significance testing doing?

Here's a little simulation where we have two groups, a treatment groups and a control group. We're going to simulate observations from both groups. We'll run the simulation two ways.

- First simulation (Null case): the treatment has no effect
- Second simulation (Non-null case): the treatment on average increases outcome

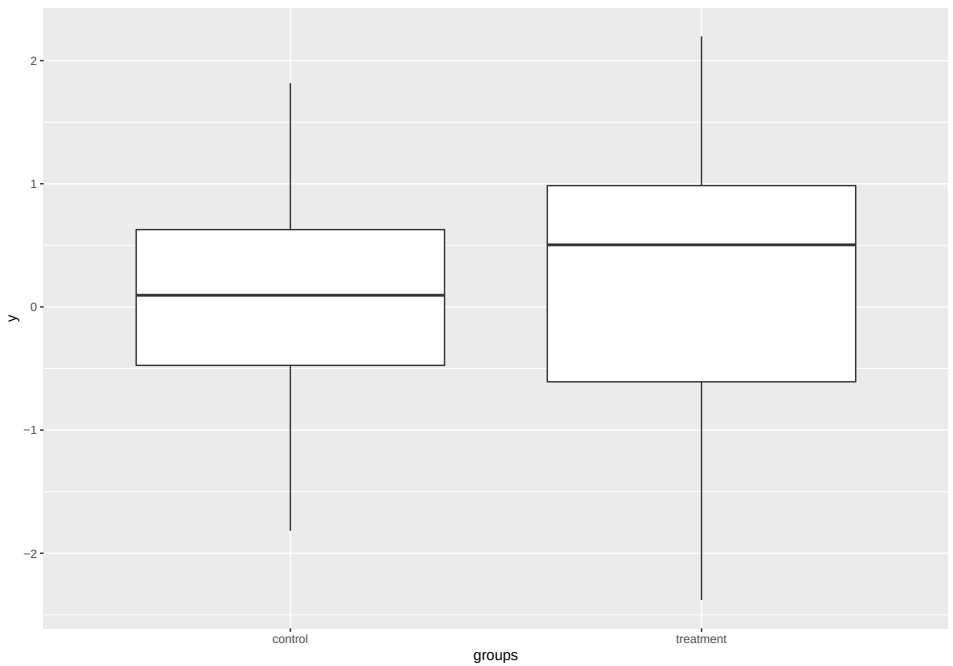
```
set.seed(12345)
# Function to generate data
generateSimulationData <- function(n1, n2, mean.shift = 0) {
  y <- rnorm(n1 + n2) + c(rep(0, n1), rep(mean.shift, n2))
  groups <- c(rep("control", n1), rep("treatment", n2))
  data.frame(y = y, groups = groups)
}
```

Let's look at a single realization in the null setting.

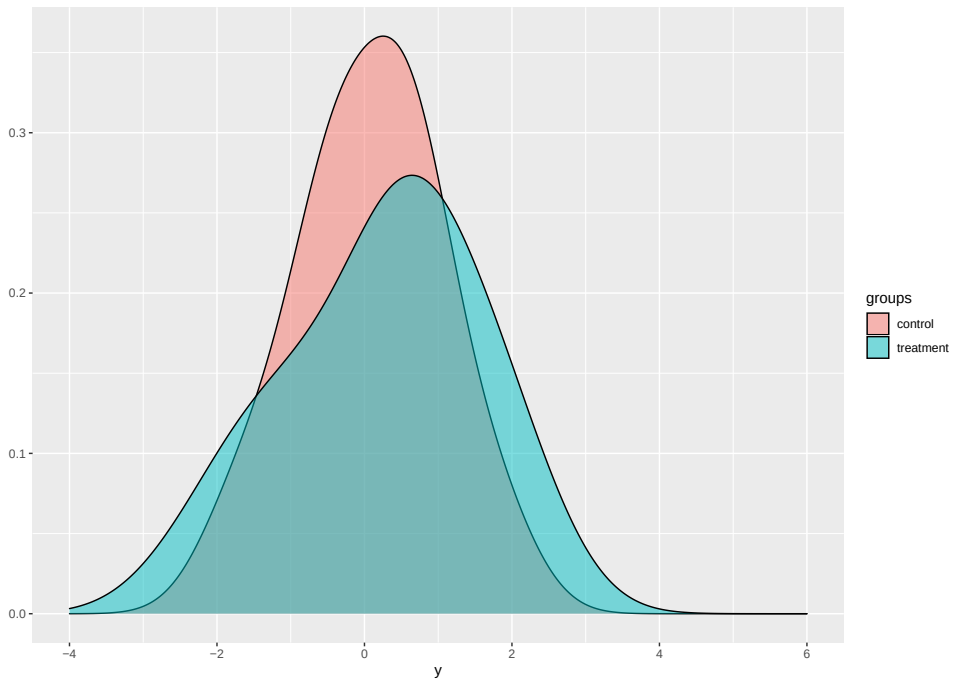
```
n1 = 30
n2 = 40
# Observation, null case
obs.data <- generateSimulationData(n1 = n1, n2 = n2)
```

```
#              y    groups
# 1   0.58552882   control
# 2   0.70946602   control
# 3  -0.10930331   control
# 4  -0.45349717   control
# 5   0.60588746   control
# 6  -1.81795597   control
# ...
# 29  0.61212349   control
# 30 -0.16231098   control
# 31  0.81187318 treatment
# 32  2.19683355 treatment
# 33  2.04919034 treatment
# 34  1.63244564 treatment
# 35  0.25427119 treatment
# ...
# 66 -1.83237731 treatment
# 67  0.88813943 treatment
# 68  1.59348847 treatment
# 69  0.51685467 treatment
# 70 -1.29567168 treatment
```

```
# Box plots
qplot(x = groups, y = y, data = obs.data, geom = "boxplot")
```



```
# Density plots
qplot(fill = groups, x = y, data = obs.data, geom = "density",
      alpha = I(0.5),
      adjust = 1.5,
      xlim = c(-4, 6))
```



```r
# t-test
t.test(y ~ groups, data = obs.data)
```

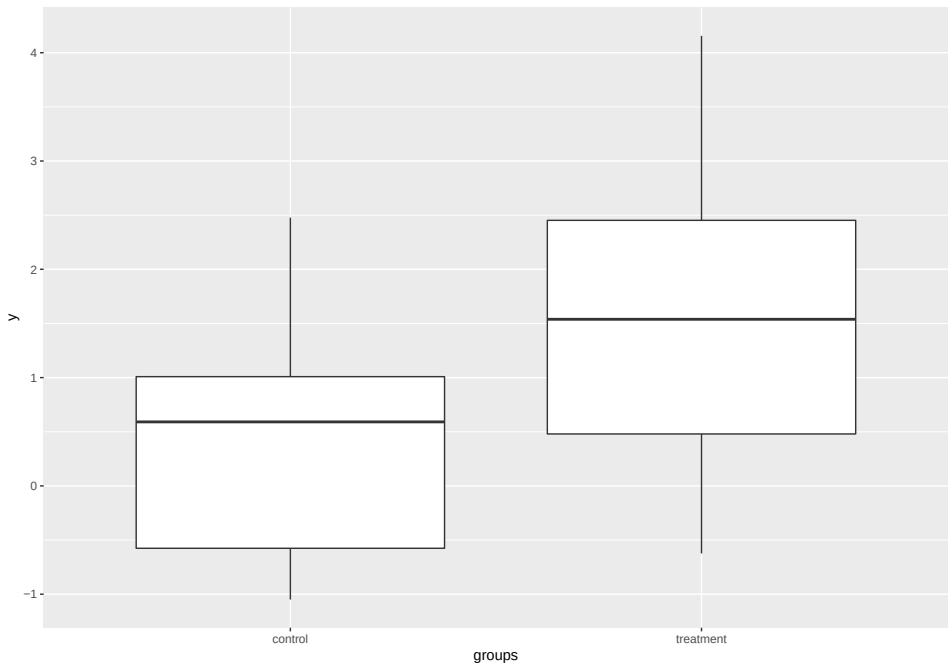
```
    Welch Two Sample t-test

data:  y by groups
t = -0.61095, df = 67.998, p-value = 0.5433
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -0.6856053  0.3641889
sample estimates:
  mean in group control mean in group treatment
              0.07880701                0.23951518
```

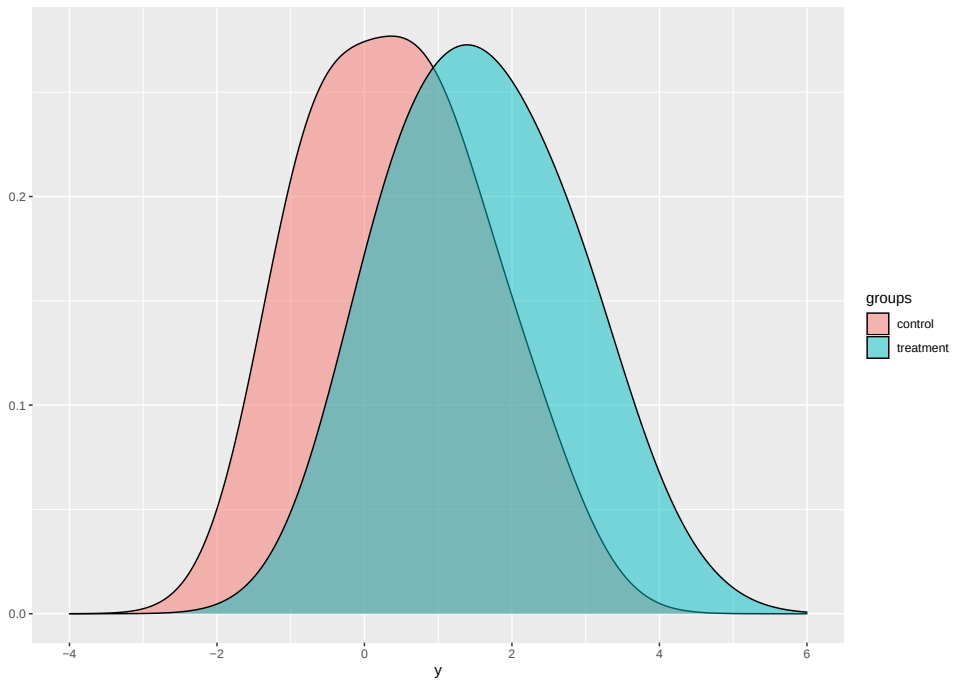
And here's what happens in a random realization in the non-null setting.

```
# Non-null case, very strong treatment effect
# Observation, null case
obs.data <- generateSimulationData(
  n1 = n1, n2 = n2, mean.shift = 1.5)

# Box plots
qplot(x = groups, y = y, data = obs.data, geom = "boxplot")
```



```
# Density plots
qplot(fill = groups, x = y, data = obs.data, geom = "density",
      alpha = I(0.5),
      adjust = 1.5,
      xlim = c(-4, 6))
```



```r
# t-test
t.test(y ~ groups, data = obs.data)
```

```
    Welch Two Sample t-test

data:  y by groups
t = -4.3081, df = 64.785, p-value = 5.708e-05
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
 -1.6911828 -0.6197985
sample estimates:
  mean in group control mean in group treatment
              0.4191634                1.5746541
```

Another way to visualize the difference

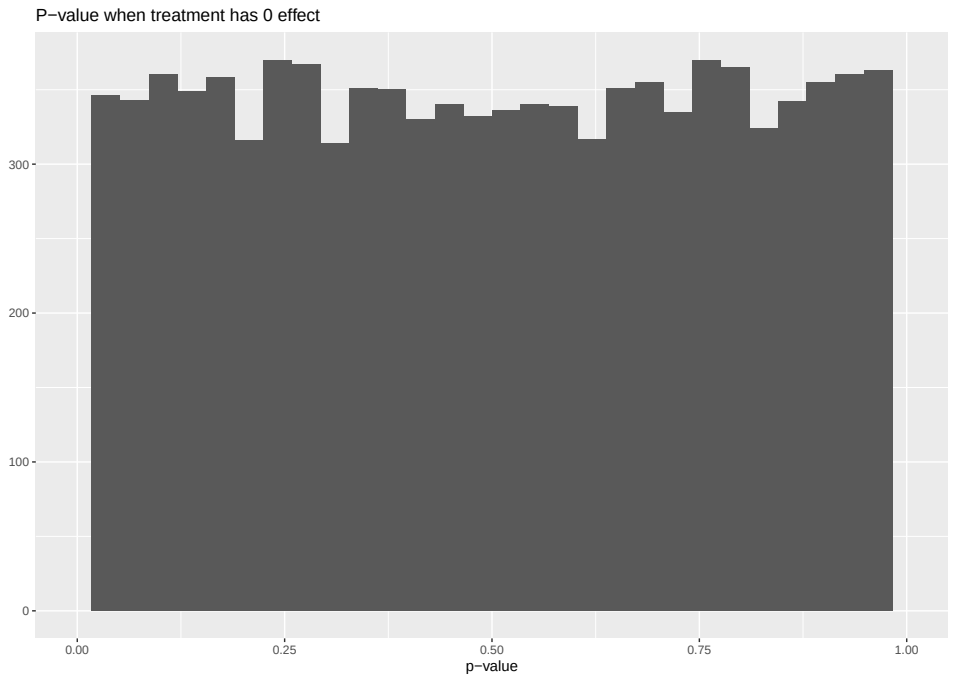
On the following few slides, we will see a few plots of p-values that illustrate what's going on behind the scenes. However, code for creating those plots is a bit beyond our course; if you are curious you can view the full original source.

More interestingly, let's see what happens if we repeat our simulation 10000 times and look at the p-values. We'll use a moderate effect of 0.5 instead of the really strong effect of 1.5 in this simulation.

Here are p-values under 0 treatment effect

```
`stat_bin()` using `bins = 30`. Pick better value with `binwic
```

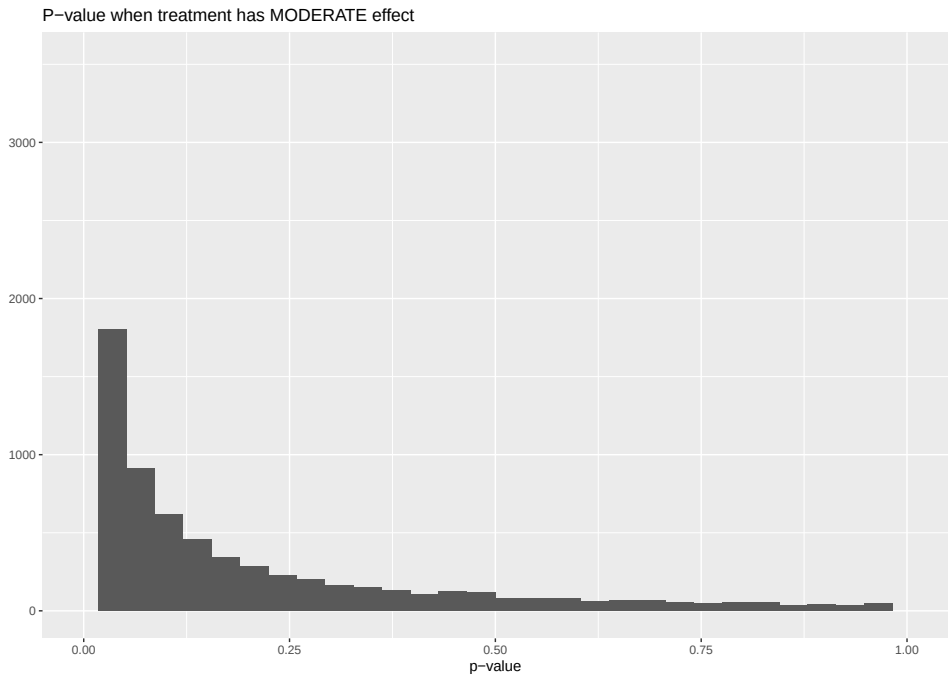
```
Warning: Removed 2 rows containing missing values (geom_bar).
```



Here are p-values under moderate treatment effect:

```
`stat_bin()` using `bins = 30`. Pick better value with `binwic
```

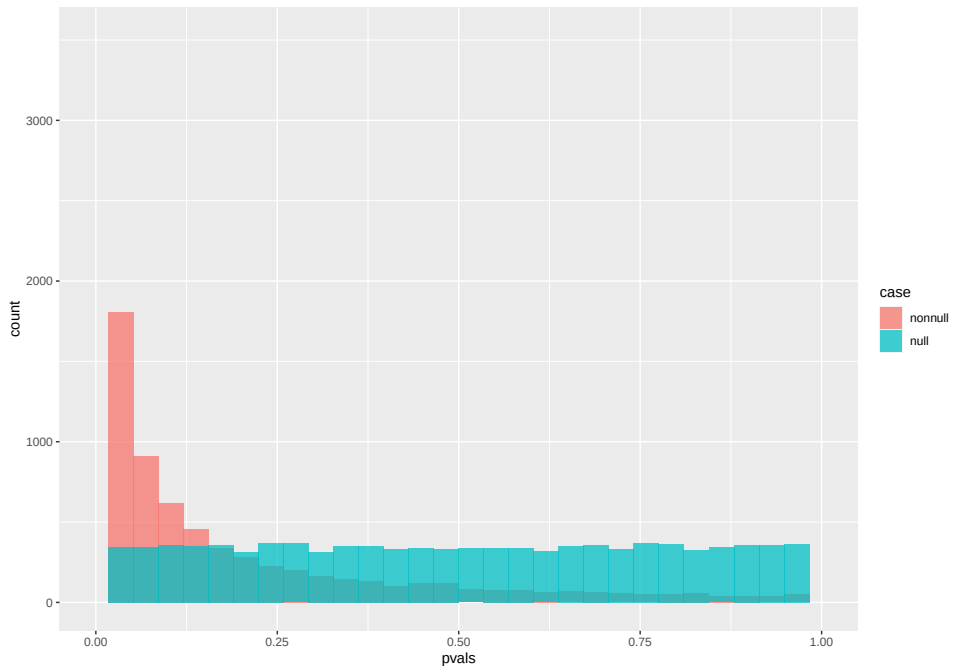
```
Warning: Removed 2 rows containing missing values (geom_bar).
```



Let's show both histograms on the same plot.

```
`stat_bin()` using `bins = 30`. Pick better value with `binwic

Warning: Removed 4 rows containing missing values (geom_bar).
```



What if sample is small and data are non-Gaussian?

- We approach t-test with caution.
  - If your data is highly skewed, you would need a very large sample size for the t-statistic to actually be t-distributed.
- When it doubt, you can run a non-parametric test.
  - This course doesn't cover this topic.

Is the data normal?

I would recommend using a non-parametric test when the data appears highly non-normal and the sample size is small. If you really want to stick to t-testing, it's good to know how to diagnose non-normality.

Remember!

The simplest thing to look at is a normal qq plot of the data. This is obtained using the `stat_qq()` function.

We have done this in a previous lecture, about 2-3 weeks ago.

License

This document is created for ITMD/ITMS/STAT 514, Spring 2021, at Illinois Tech.

The simulated example part of this lecture and the contingency table example part of this lecture were sourced from Prof. Alexandra Chouldechova, released under a Attribution-NonCommercial-ShareAlike 3.0 United States license.