Nonparametric and Simulation-Based Tests

STAT 3202 @ OSU, Spring 2019

Dalpiaz

What is **Parametric** Testing?

Warmup #1, Two Sample Test for $p_1 - p_2$

Ohio Issue 1, the **Drug and Criminal Justice Policies Initiative**, is on the ballot in Ohio as an initiated constitutional amendment on November 6, 2018. Among other things, this amendment seeks to make offenses related to drug possession and use no more than misdemeanors.

Suppose some pollster obtains random samples of registered Democrats and Republicans:

- Democrats: $n_D = 100$, 60 supporters
- Republicans: $n_R = 150$, 60 supporters

Use this data to test $H_0: \ p_D = p_R$ vs $H_1: \ p_D \neq p_R$ where p_D is the proportion of Democrats that support this issue.

Report:

- The **test statistic**
- The **p-value**
- A **decision** when $\alpha = 0.01$.

Warmup #2, Paired Sample Test

- Data from 1993 article (BMJ, Scanlon et al.) *"Is Friday the 13th bad for your health?"*
- Researchers counted the **number of emergency admissions** due to transportation accidents at South West Thames Regional Hospital Authority on **six pairs of consecutive Fridays** – a Friday the 6^{th} and a Friday the 13^{th} in 1989-1992
- Use the following data to test H_0: $\mu_{13} = \mu_6$ vs H_1: $\mu_{13} > \mu_6$. Use $\alpha = 0.05$.

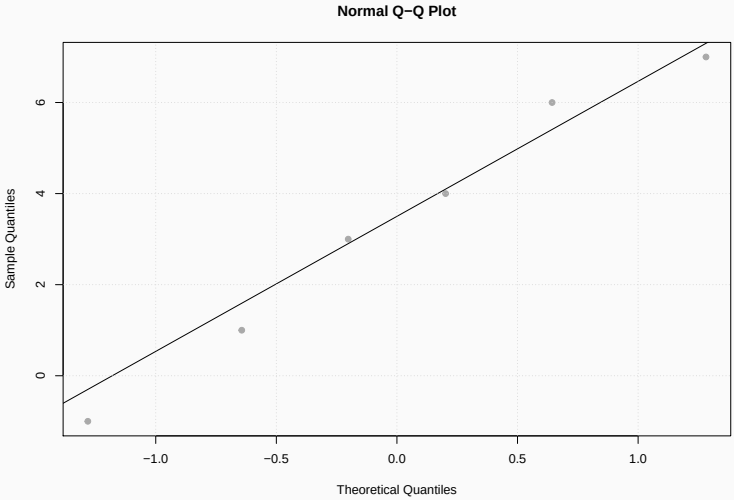
```
##   year     month Friday_6 Friday_13
## 1 1989   October        9        13
## 2 1990      July        6        12
## 3 1991 September       11        14
## 4 1991  December       11        10
## 5 1992     March        3         4
## 6 1992  November        5        12
```

Warmup #2, Difference Data

```
##   year     month Friday_6 Friday_13 diff
## 1 1989   October        9         13    4
## 2 1990      July        6         12    6
## 3 1991 September       11         14    3
## 4 1991  December       11         10   -1
## 5 1992     March        3          4    1
## 6 1992  November        5         12    7

##   mean_d     sd_d
## 3.333333 3.011091
```

Warmup #2, A Note on Assumptions



Warmup #3, Two Sample Test for $\mu_1 - \mu_2$

Suppose a researcher is interested in the effects of a vegetarian diet on health. They obtain random samples of 15 adult female vegetarians and 10 adult female omnivores.

The vegetarians have a sample mean weight of 55 kilograms with a sample standard deviation of 5 kilograms. The omnivores have a sample mean weight of 60 kilograms with a sample standard deviation of 6 kilograms.

Use this data to test $H_0: \ \mu_V = \mu_O$ vs $H_1: \ \mu_V \neq \mu_O$. Use $\alpha = 0.05$

Nonparametric versus Parametric Methods

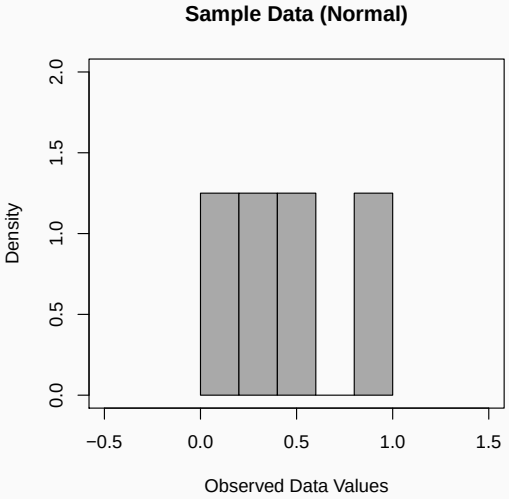
- **Parametric** Testing Methods
  - Methods that make distribution assumptions about the data up to a finite number of values – the **parameters**
  - e.g. the one-sample t-test assumes: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$
    - parameters μ and σ unknown
  - Can also be applied more generally by invoking robustness and large sample properties
- **Nonarametric** Testing Methods
  - Anything that is **not** parametric
  - e.g. $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ population with median m
    - no other assumptions!
  - e.g. $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ population with a symmetric distribution
    - no other assumptions!

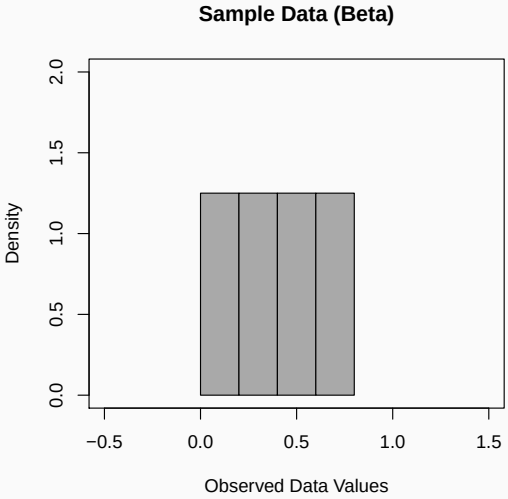
What Makes a Test Valid?

Question: Do we feel comfortable applying a one-sample t-test of $H_0 : \mu = 0.5$ to either of these datasets? Is the one-sample t-test valid?

```
set.seed(2)
sample_norm = rnorm(n = 4, mean = 0.5, sd = 1 / sqrt(12))
sample_unif = rbeta(4, shape1 = 1 / 3, shape2 = 1 / 3)
```

“Small” Sample Data, $n = 4$





Checking Validity (Normal Case)

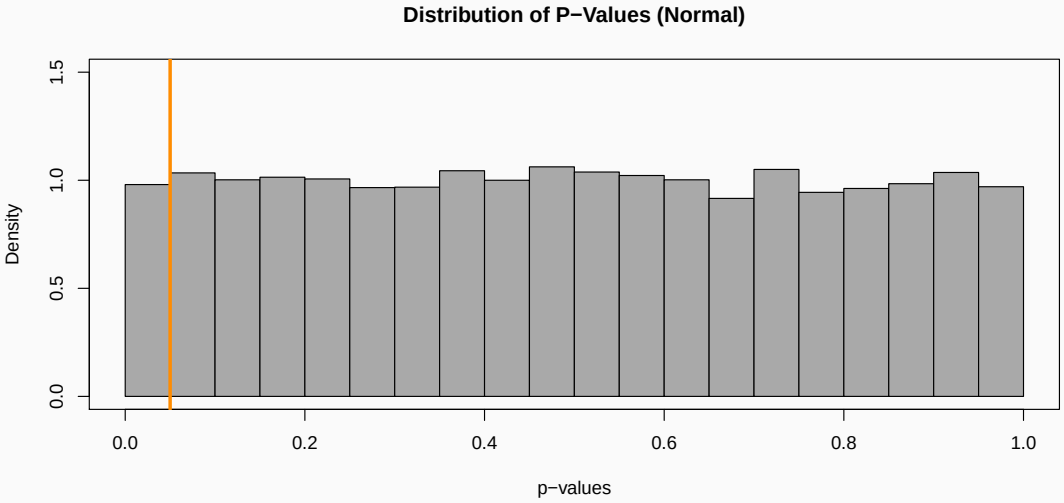
- A test is valid if the actual Type I Error rate is the claimed α level.
  - If we run the a test using $\alpha = 0.05$ over and over and H_0 is true, we reject H_0 (no more than) 5% of the time. (Check with simulation!)
    - If we reject roughly 5% of the time, the test is **valid**.
    - If we reject less than 5% of the time, the test is **conservative**, but still "valid."
    - If we reject more than 5% of the time, the test is **invalid** and should not be used.

```
set.seed(42)
p_vals_norm = replicate(n = 10000,
    t.test(rnorm(n = 4, mean = 0.5, sd = 1 / sqrt(12)), mu = 0.5)$p.value
)
```

```
mean(p_vals_norm < 0.05)
```

```
## [1] 0.049
```

A Valid Testing Example, Normal



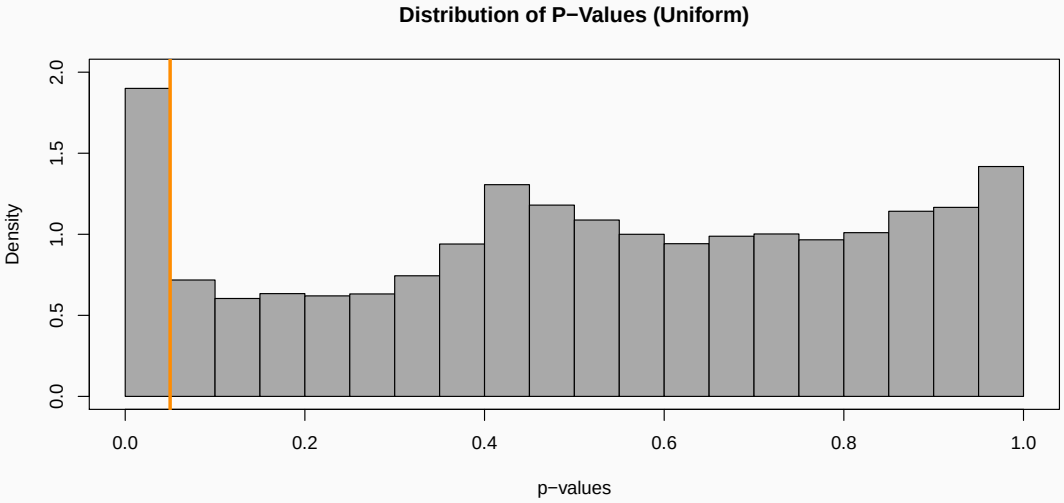
An Invalid Testing Example, Uniform

```
set.seed(42)
p_vals_unif = replicate(n = 10000,
    t.test(rbeta(4, shape1 = 1 / 3, shape2 = 1 / 3), mu = 0.5)$p.value
)
```

```
mean(p_vals_unif < 0.05)
```

```
## [1] 0.095
```

An Invalid Testing Example, Uniform

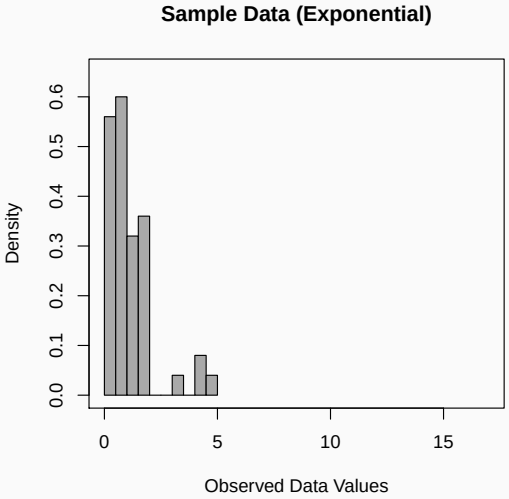


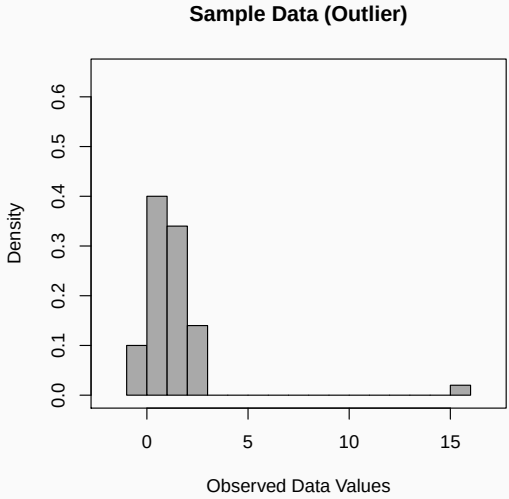
Is a Test Valid?

Question: Do we feel comfortable applying a one-sample t-test of $H_0 : \mu = 1$ to either of these datasets? Is the one-sample t-test valid?

```
set.seed(2)
sample_exp = rexp(n = 50, rate = 1)
sample_out = c(rnorm(n = 49, mean = 1), rnorm(n = 1, mean = 15))
```

Large Sample Data, Non-Normal and Outlier





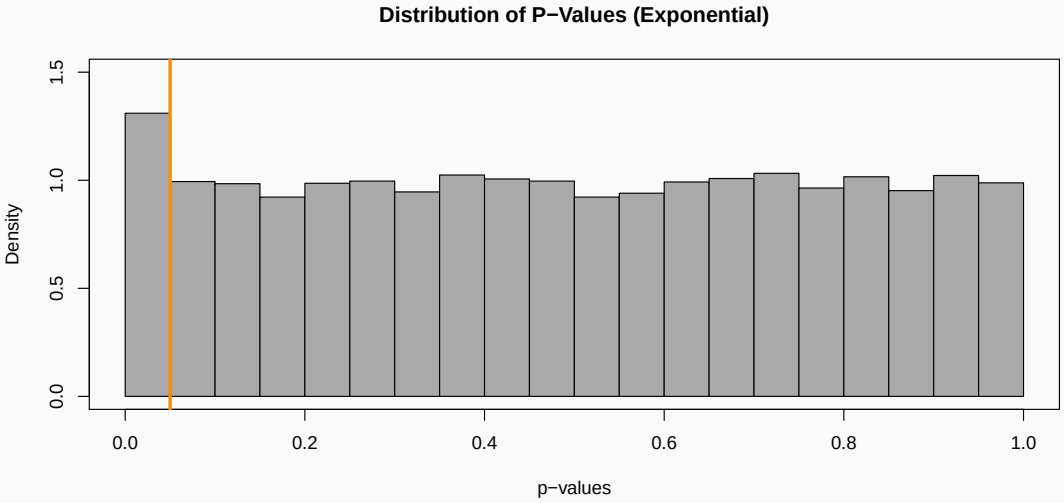
Simulation Study, Exponential

```
set.seed(42)
p_vals_exp = replicate(n = 10000,
    t.test(rexp(n = 50, rate = 1), mu = 1)$p.value
)
```

```
mean(p_vals_exp < 0.05)
```

```
## [1] 0.0655
```

Simulation Study, Exponential



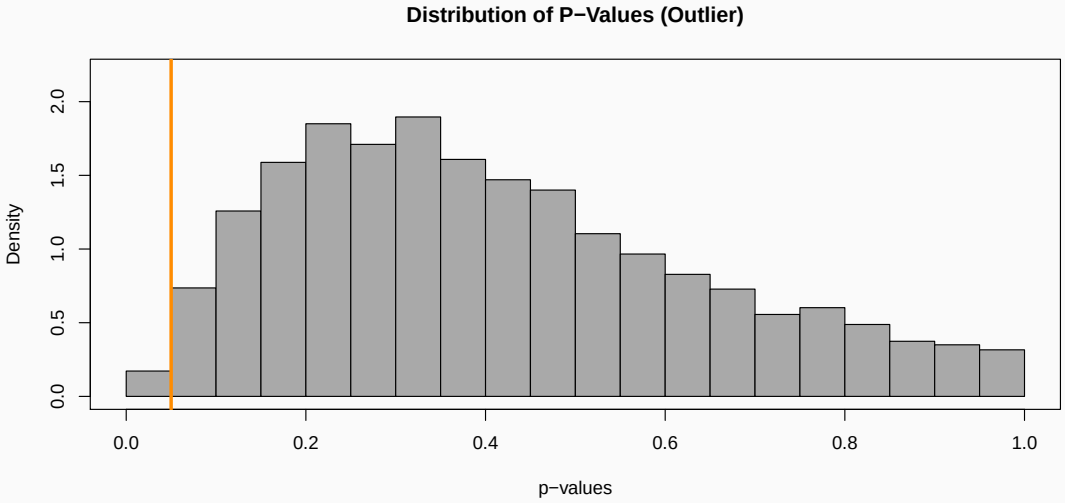
Simulation Study, Outlier

```
set.seed(42)
p_vals_out = replicate(n = 10000,
    t.test(c(rnorm(n = 49, mean = 1), rnorm(n = 1, mean = 15)), mu = 1)$p.value
)
```

```
mean(p_vals_out < 0.05)
```

```
## [1] 0.0086
```

Simulation Study, Outlier

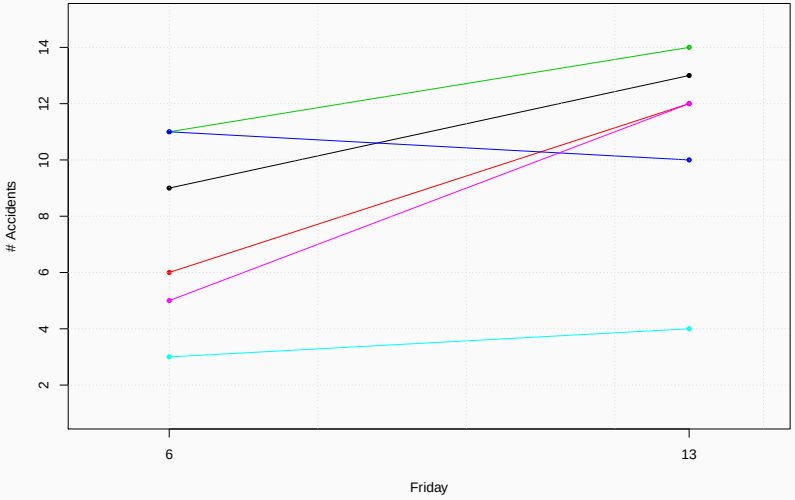


Friday the 13th

- Data from 1993 article (BMJ, Scanlon et al.) *"Is Friday the 13th bad for your health?"*
- Researchers counted the **number of emergency admissions** due to transportation accidents at South West Thames Regional Hospital Authority on **six pairs of consecutive Fridays** – a Friday the 6^{th} and a Friday the 13^{th} in 1989-1992
- The data:

```
##   year     month Friday_6 Friday_13 diff
## 1 1989   October        9        13    4
## 2 1990      July        6        12    6
## 3 1991 September       11        14    3
## 4 1991  December       11        10   -1
## 5 1992     March        3         4    1
## 6 1992  November        5        12    7
```

Friday the 13th



Example: Friday the 13th

- Researchers were interested in determining whether accident rates **tend to be higher** on Friday the 13ths compared with other Fridays, as exemplified by Friday the 6ths
- Define appropriate parameters and state the null and alternative hypotheses

- Should we use procedures for independent data or procedures for matched data?

Possible Analyses

- The "paired" or "matched" t-test: take the **difference** between the number of accidents on the paired Fridays; check the assumption that the **difference** may plausibly come from a normal distribution; run a 1-sample t-test
- The Sign Test [new!]
- Wilcoxon Signed-Rank Test [new!]
- A Permutation Test [new!]

Why **Nonparametric** Testing?

Nonparametric Testing

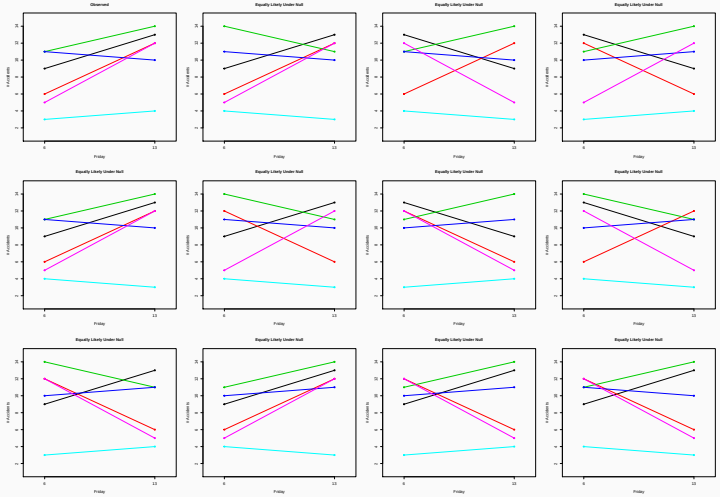
Is useful when...

- the sample size is very small
- the distributional assumptions of a parametric test are doubtful (especially in the presence of outliers)
- when the variable of interest is ordinal
  - e.g., bakers bake pies (with butter crust and with lard crust) and judges eat pieces and give each pie a number of stars (from 1 to 4).
  - treating these scores as strictly quantitative may not make sense (e.g., is the difference between a 2 and a 3 "the same" as the difference between a 3 and a 4?)
  - nonparametric tests exist to answer the question "are butter crusts tastier than lard crusts?" that rely on the ranking of the pies but not the absolute value of the score

The Sign Test

```
##   year     month Friday_6 Friday_13 diff
## 1 1989   October        9         13    4
## 2 1990      July        6         12    6
## 3 1991 September       11         14    3
## 4 1991  December       11         10   -1
## 5 1992     March        3          4    1
## 6 1992  November        5         12    7
```

Permutation Testing



Two-Sample Example: Desensitization to Violence

- Molitor (1989) conducted a study to see if children who watched TV or film violence were significantly more (or less) tolerant of "real-life" violence, versus children watching a nonviolent program.
- Half of 42 children were shown violent TV (an edited version of *The Karate Kid*).
- The other half watched exciting but nonviolent sports (highlights from the 1984 Summer Olympics).
- Each child was asked to "watch over" two younger children in the next room via a television monitor, and go and get the research assistant if the younger children "got into trouble."
- Each child actually witnessed a videotaped sequence depicting two small children first playing with blocks, then progressively get more violent – pushing each other, chasing each other, fighting, knocking down a video camera while fighting.

Two-Sample Example: Desensitization to Violence

- Toleration of violence was measured by the time (in seconds) each child stayed in the room after he or she witnessed the two younger children's first act of violence.
- (Each child was subsequently assured that an adult had entered the room, the children were not hurt, and the video camera was not damaged.)

```
karate_kid = c(37, 39, 30, 7, 13, 139, 45, 25, 16, 146, 94, 16, 23, 1, 290, 169,
               62, 145, 36, 20, 13)
olympics = c(12, 44, 34, 14, 9, 19, 156, 23, 13, 11, 47, 26, 14, 33, 15, 62, 5,
             8, 0, 154, 146)
```