HYPOTHESIS TESTING USING R, EXPLORATORY DATA ANALYSIS

Jack Hester, MPH

CEPC 0904

Summer 2022

REVIEW: HYPOTHESIS TESTING

- Generate hypotheses
- Test them with the appropriate test
- Usually you will have a null (H_0) and Alternative (H_A)
- There are general research questions/hypotheses, there are also specific hypotheses depending on the statistical method

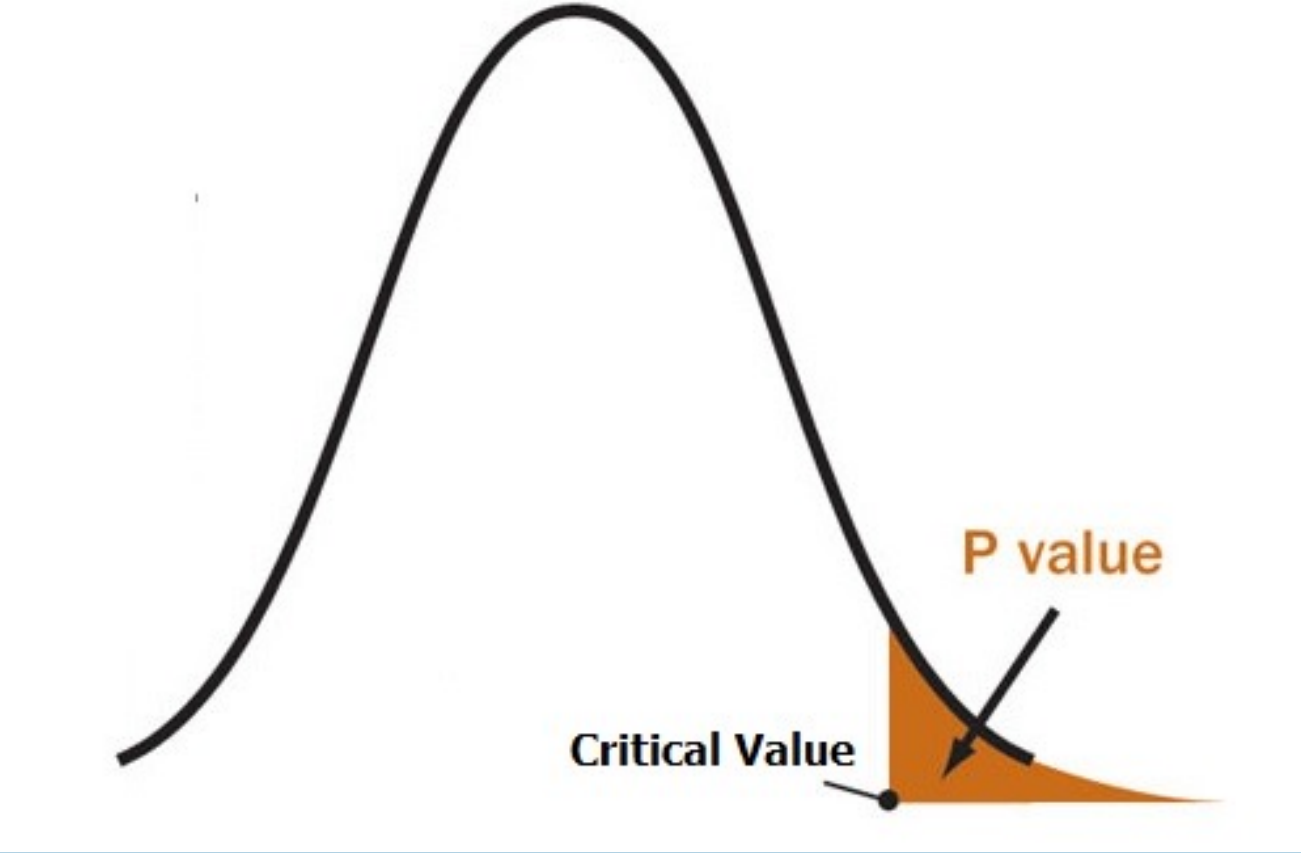
TERMINOLOGY

- Significance level (α) = "probability of rejecting a null hypothesis when it's true"
  - Loosely this means you don't pick up on a valid, positive result
- Distribution = representation of probabilities of where a value will lie
- p-value (p) = the "probability of obtaining a result at least as extreme as the observed result assuming the null hypothesis is correct"
  - In other words, probability of getting that significant result by random chance alone

TERMINOLOGY CONT.

- Critical value = cutoff value in distribution, calculated depending on significance level
  - Use value table or calculator, for sig. level of 0.05 the critical value for a 2-sample test is 1.96
- Confidence interval (CI) = A range of values (interval) in which you expect the true population mean to lie
- Quantile = % of values below a certain value

EXAMPLE GRAPHIC (ONE-SIDED TEST)



https://www.analyticsvidhya.com/blog/2020/11/interpreting-p-value-and-r-squared-score-on-real-time-data-statistical-data-exploration/

TESTS IN R

Z TEST COMMANDS

```
z.test(
  x,
  y = NULL,
  alternative = "two.sided",
  mu = 0,
  sigma.x = NULL,
  sigma.y = NULL,
  conf.level = 0.95
)
```

sample

Second sample, NULL if one sample

one or two sided

Mean to test against if one-sample

Population standard deviation of sample 1 if known

Population standard deviation of sample 2 if known

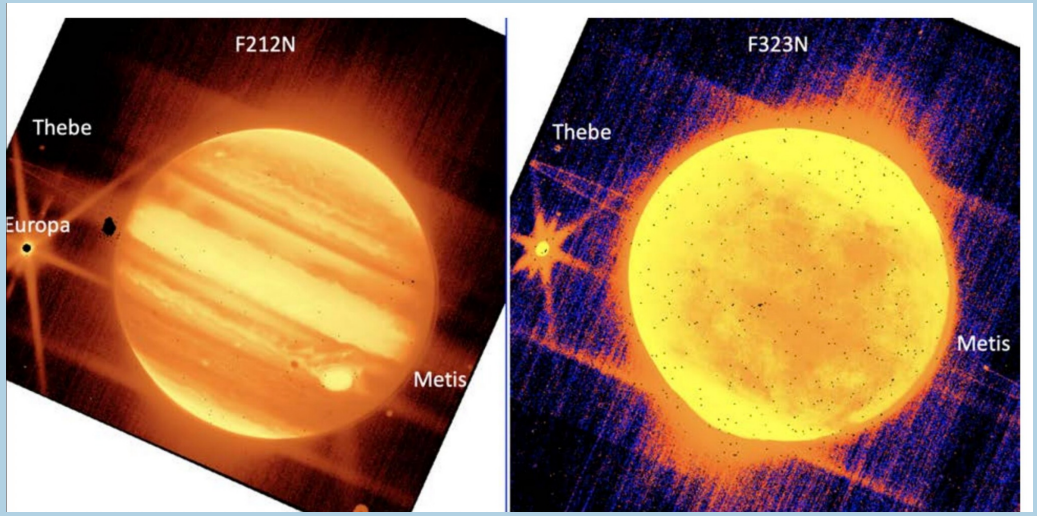
Confidence interval size

T TEST COMMANDS

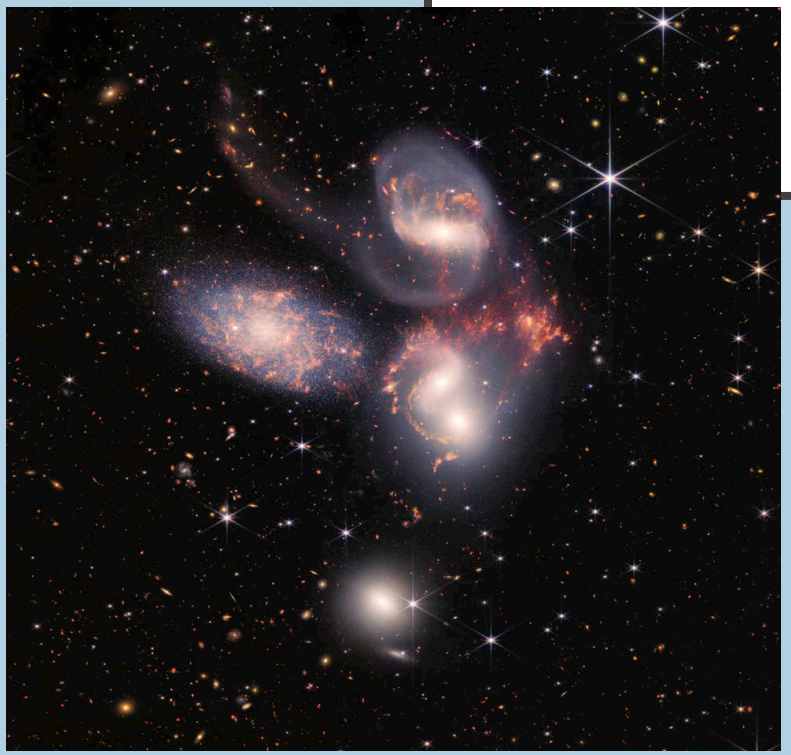
```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, …)
```

Two sided or one sided, if one sided less or greater than the mean depending on what we're interested in testing

Variance equal or not depends on if we know the samples' variances to be meaningfully different (can look this up)



BREAK



R EXAMPLE – 2 SAMPLE T-TEST

```
> t.test(s1, s2, paired=FALSE)

        Welch Two Sample t-test

data:  s1 and s2
t = -5.3982, df = 23.522, p-value = 1.624e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.901571  -7.098429
sample estimates:
mean of x mean of y
      5.1      16.6
```

CI doesn't cover 0, p << 0.05 so significant result

We are 95% confident that the interval (-15.9, -7.1) contains the mean difference

R EXAMPLE – 2 SAMPLE T-TEST

```
> t.test(s1, s2, paired=FALSE)

        Welch Two Sample t-test

data:  s1 and s2
t = -5.3982, df = 23.522, p-value = 1.624e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.901571  -7.098429
sample estimates:
mean of x mean of y
      5.1      16.6
```

What are we testing? Assume s1 and s2 represent number of youtube videos watched by students at two universities

REFERENCE GUIDE

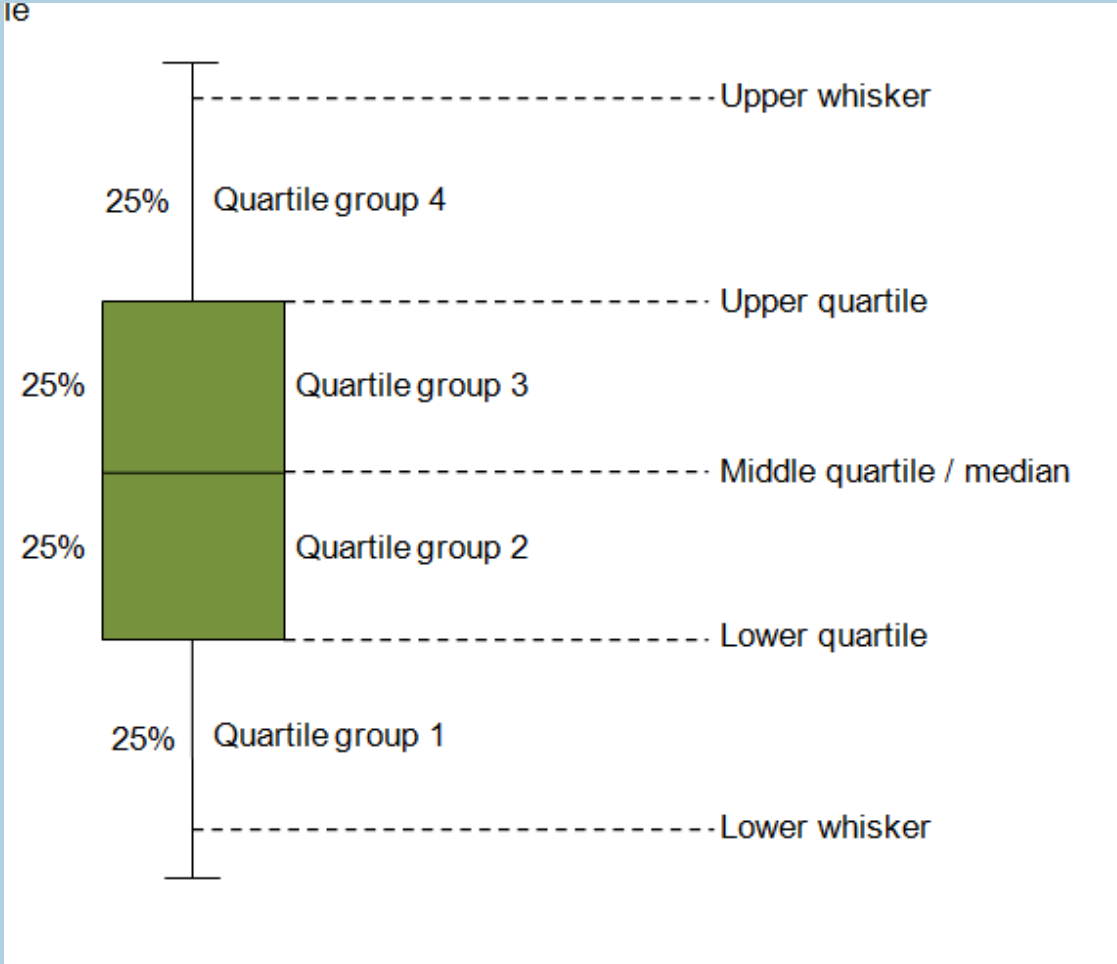
- https://cepc0904-22.jackhester.com/documents/stats_guide.pdf

EXPLORATORY DATA ANALYSIS (EDA) TOOLS

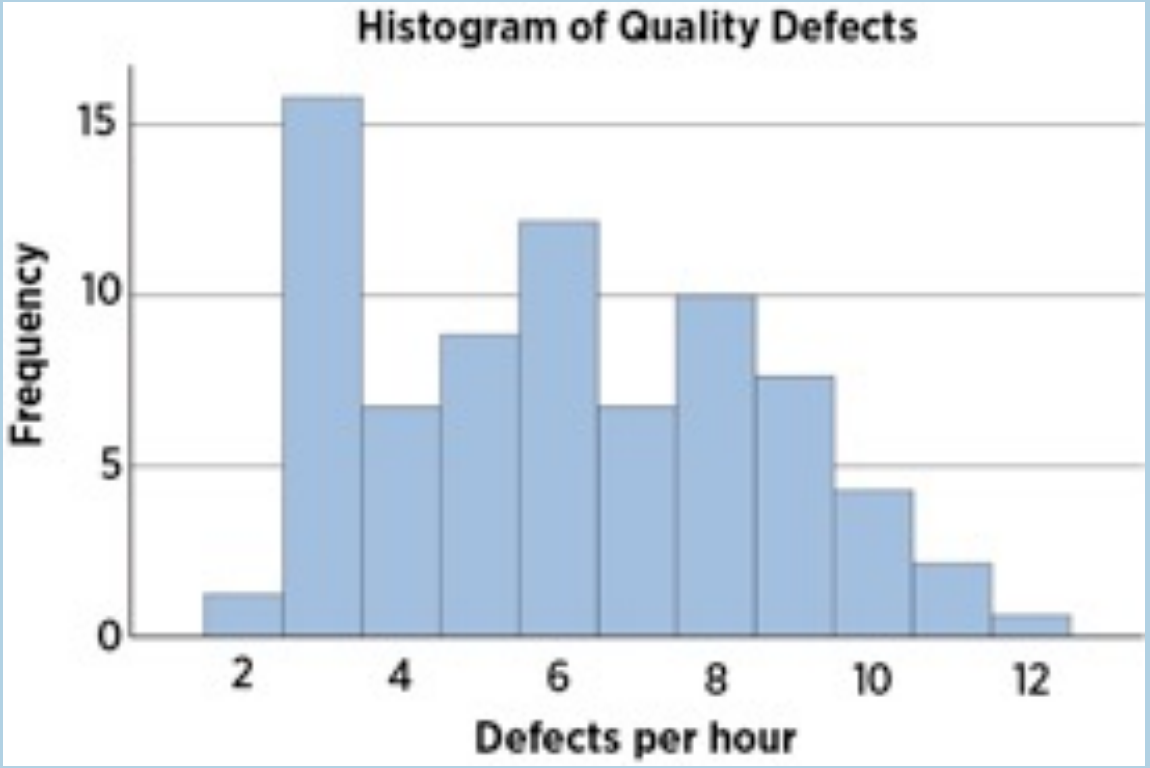
APPLICATIONS

- Understanding data
- Checking assumptions of tests
- Looking for interesting outliers

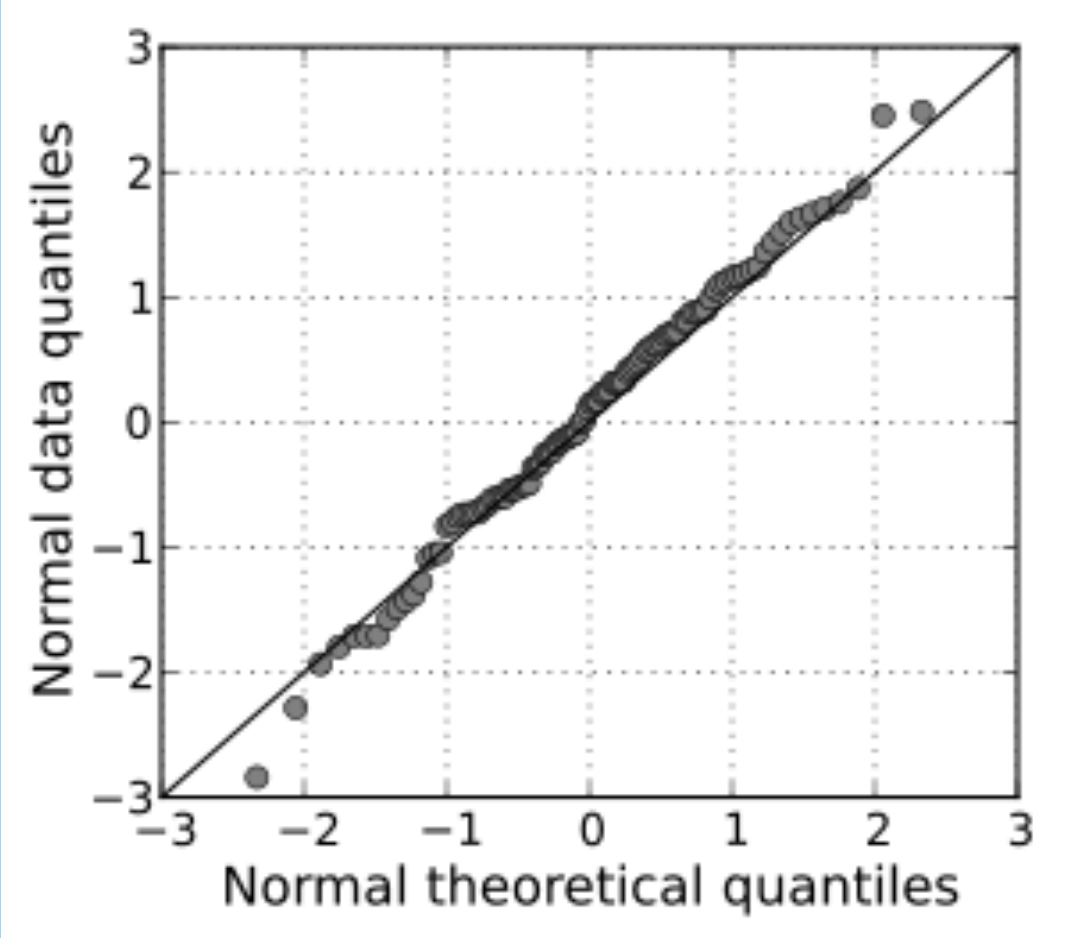
BOXPLOTS


Upper whisker
25%
Quartile group 4
Upper quartile
25%
Quartile group 3
Middle quartile / median
25%
Quartile group 2
Lower quartile
25%
Quartile group 1
Lower whisker

HISTOGRAMS


Histogram of Quality Defects
Frequency
15
10
5
0
2
4
6
8
10
12
Defects per hour

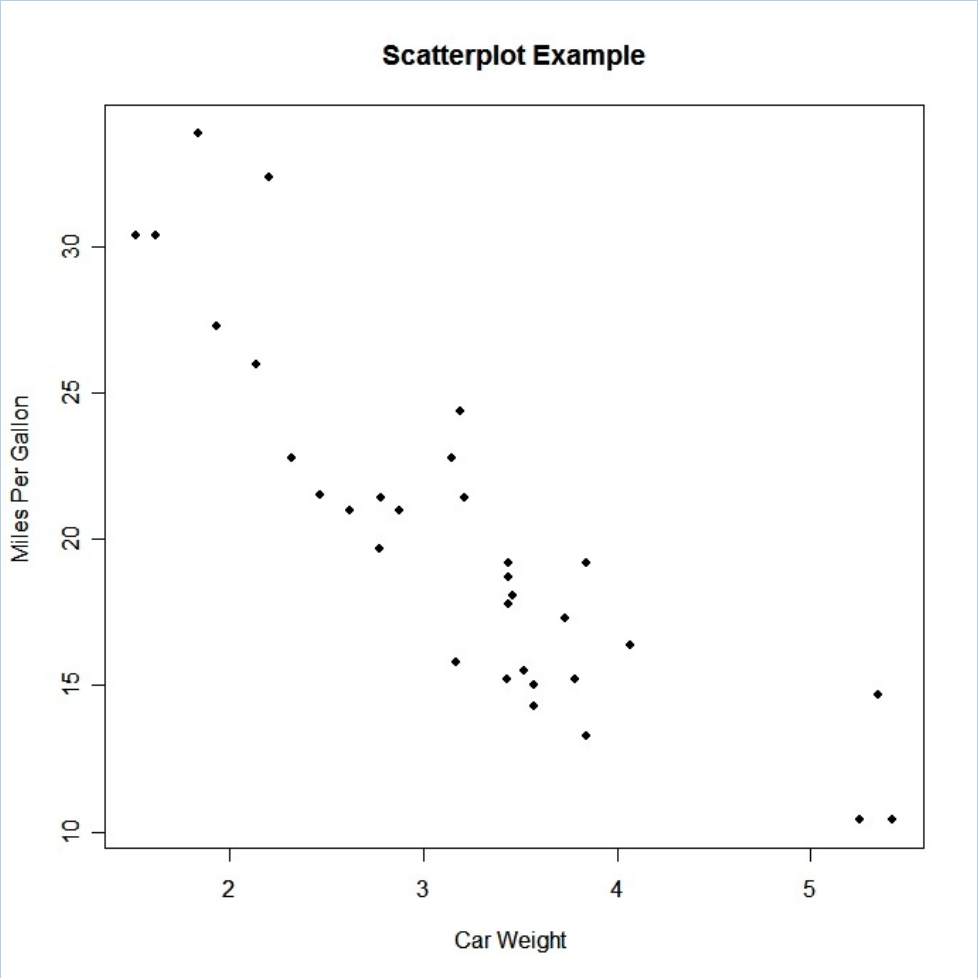
Q-Q PLOT (Q=QUANTILE)



MAIN EDA STEP

- Explore this ratio

SCATTERPLOT



NEXT STEPS

- Watch probability video (tomorrow)
- Work on HW 5 (project assignment) due Wednesday
- Work on HW 4 due Friday (groups released later today)
- Review stats guide (under lecture notes today/last week)
- "Statistics for Non-statisticians" available on Canvas
  - https://canvas.brown.edu/courses/1088637/files/folder/unfiled?preview=68369965