Statistical Computing and Data Visualization in R

Lecture 4

Basic statistical analysis

Statistical Analysis in R

- R provides a large number of functions that implement widely (and not so widely!) used statistical methods and streamline their application.
- We will cover some of those tools, particularly those related simple tests for means, variances and proportions, and for linear regression

Tests of proportions

- Consider situations where data comes from binomial distributions: Your data *x* is the number of successes out of *n* independent and identically distributed trials.
- Examples:
  - ***x*** = number of patients in the emergency room that suffer a particular complication, ***n*** = total number of patients coming in the emergency.
  - ***x*** = number of respondents in survey who think POTUS is doing a great job, ***n*** = total number of people surveyed.

Tests of proportions

- In this case $x \sim \text{Binomial}(n, \theta)$, where n is the known number of trials and θ is the unknown probability of success in each individual trial.
- The "best" estimator for θ is the observed proportion of successes x/n.
- We often want to test
  - H0: $\theta = \theta_\theta$ vs H1: $\theta \neq \theta_\theta$ (or $\theta > \theta_\theta$ or $\theta < \theta_\theta$) ➔ One sample test, θ is unknown but θ_θ is a fixed predetermined value)
  - H0: $\theta_1 = \theta_2$ vs H1: $\theta_1 \neq \theta_2$ (or $\theta_1 > \theta_2$ or $\theta_1 < \theta_2$) ➔ Two sample test, both θ_1 and θ_2 are unknown and we have samples (x_1, n_1) and (x_2, n_2).

Tests of proportions

- Examples:
  - Is the proportion of defective pieces being produced by the factory greater than the maximum of θ_θ = 3% specified in the contract?
  - Do UCSC and UCSB admit Hispanic students at the same rate?
  - Is a new needle exchange program effective at reducing the rate of Hepatitis C infection?

Tests of proportions

- For the one-sample test, under H0:

$$\frac{\hat{x} - \theta_0}{\sqrt{\hat{x}(1-\hat{x})/n}} \sim N(0,1) \qquad \hat{x} = \frac{x}{n}, \quad \text{"large" } n$$

- For the two-sample test, under H0:

$$\frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{\hat{x}_1(1-\hat{x}_1)}{n_1} + \frac{\hat{x}_2(1-\hat{x}_2)}{n_2}}} \sim N(0,1)$$

There are a few different variants, which differ in the denominator

$$\hat{x}_1 = \frac{x_1}{n_1}, \quad \hat{x}_2 = \frac{x_2}{n_2}, \quad \text{"large" } n_1 \text{ and } n_2$$

Tests of proportions

- Implementing one- and two-sample approximate tests of proportions:.

```
> n = c(120, 140)
> x = rbinom(2, n, 1/3)
> prop.test(x,n,alternative="greater")
    2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 0.398, df = 1, p-value = 0.2641
alternative hypothesis: greater
95 percent confidence interval:
 -0.05667545  1.00000000
sample estimates:
   prop 1    prop 2
0.3000000 0.2571429
> prop.test(x[1],n[1],1/2)
    1-sample proportions test with continuity correction

data:  x[1] out of n[1], null probability 1/2
X-squared = 18.408, df = 1, p-value = 1.783e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2215321 0.3914945
sample estimates:
  p
0.3
```

Tests of proportions

- The function `prop.test()` generates an object with a number of attributes. This object can be stored in a variable

```
> n = c(120, 140)
> x = rbinom(2, n, 1/3)
> z = prop.test(x,n,alternative="greater")
> names(z)
[1] "statistic"   "parameter"   "p.value"     "estimate"    "null.value"
[6] "conf.int"    "alternative" "method"      "data.name"
```

- You can specifically interrogate the value of any of the components of the object:

```
> z$estimate
   prop 1    prop 2
0.3250000 0.3785714
> z$pvalue
[1] 0.7794991
> z$null.value
NULL
> z$method
[1] "2-sample test for equality of proportions with continuity
correction"
> z$null.value
NULL
> z$conf.int
[1] -0.1587387  1.0000000
attr(,"conf.level")
[1] 0.95
```

Can you explain this?

Tests of proportions

- For one-sample situations, an exact test is also implemented

```
> prop.test(x[1],n[1],1/3)
    1-sample proportions test with continuity correction

data:  x[1] out of n[1], null probability 1/3
X-squared = 0.45937, df = 1, p-value = 0.4979
alternative hypothesis: true p is not equal to 0.3333333
95 percent confidence interval:
 0.2215321 0.3914945
sample estimates:
  p
0.3
> binom.test(x[1],n[1],1/3)
    Exact binomial test

data:  x[1] and n[1]
number of successes = 36, number of trials = 120, p-value = 0.4981
alternative hypothesis: true probability of success is not equal to
0.3333333
95 percent confidence interval:
 0.2197565 0.3903961
sample estimates:
probability of success
                   0.3
```

- Repeat with n small and check the difference in p-values.

Tests of means

- Consider now the same situation, but where measurements correspond to continuous values (height, temperature, weight)
- We again assume that observations are collected independently, but now we assume that they are coming from a normal distribution with unknown mean(s) and variance(s).

Tests of means

- Again, we have two cases:
  - **One sample test:** Is the average concentration of a contaminant in n = 10 samples below the <u>known</u> EPA standard?

    $x_i \sim N(\mu, \sigma^2)$ H0: $\mu = \mu_0$ vs H1: $\mu \neq \mu_0$
  - **Two sample test:** Is the average yield per acre of maize the same under the new fertilizer as it was under the old one (<u>both estimated from data</u>)?

    $x_i \sim N(\mu_x, \sigma_x^2)$ $y_i \sim N(\mu_y, \sigma_y^2)$

    H0: $\mu_x = \mu_y$ vs H1: $\mu_x > \mu_y$

Tests of means

- For the one-sample test, under H0:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0,1) \qquad \bar{x} = \frac{x}{n}, \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- For the two-sample test, under H0:

$$\frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim N(0,1)$$

Assumes equal variances (i.e., $\sigma_x = \sigma_y$)
Other variants for $\sigma_x \neq \sigma_y$ and paired tests

$$\bar{x} = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i, \quad \bar{y} = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i, \quad s^2 = \frac{\sum_{i=1}^{n_x}(x_i - \bar{x})^2 + \sum_{i=1}^{n_x}(y_i - \bar{y})^2}{n_x + n_y - 2}$$

Tests of means

- Similarly, you can easily conduct one-sample and two-sample tests of means for data from a normal distribution:

```
> x = rnorm(56, 5, 2)
> y = rnorm(74, 4.5, 1.5)
> t.test(x, y)
    Welch Two Sample t-test

data:  x and y
t = 1.4501, df = 83.526, p-value = 0.07539
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.07518682         Inf
sample estimates:
mean of x mean of y
 4.937603  4.426350
> t.test (x, mu=4.85)
    One Sample t-test

data:  x
t = 0.27939, df = 55, p-value = 0.781
alternative hypothesis: true mean is not equal to 4.85
95 percent confidence interval:
 4.309221 5.565985
sample estimates:
mean of x
 4.937603
```

Tests of means

- You can carry out tests equal-variance t-tests, as well as paired t-tests:

```
> t.test(x, y, var.equal =TRUE, alternative="greater")
    Two Sample t-test

data:  x and y
t = 0.48736, df = 128, p-value = 0.3134
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.4051416        Inf
sample estimates:
mean of x mean of y
 5.068840  4.900004
> x = rnorm(60, 5, 2)
> y = rnorm(60, 4.5, 1.5)
> t.test (x, y, paired=TRUE)
    Paired t-test

data:  x and y
t = 2.3085, df = 59, p-value = 0.02449
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1042614 1.4610188
sample estimates:
mean of the differences
              0.7826401
```

Formulas

- The function `t.test()` can also accept formula objects:

```
> x = rnorm(56, 5, 2)
> y = rnorm(74, 4.5, 1.5)
> z = c(x,y)
> g = factor(c(rep(1,length(x)), rep(2,length(y))))
> t.test(z ~ g)
 Welch Two Sample t-test

data:  z by g
t = 2.3177, df = 100.17, p-value = 0.0225
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 0.1000213 1.2890438
sample estimates:
mean in group 1 mean in group 2
       5.075129        4.380596
```

- The symbol ~ should be read as "explained by".
- Formulas are also useful for streamlined plotting (try `plot(z ~ g)` and compare the output with that of `boxplot(x, y)`).

Test of variances

- You can carry out similar tests for the variances instead of the means (this is a less frequent problem)
  - One sample: H0: $\sigma = \sigma_0$ vs H0: $\sigma \neq \sigma_0$
  - Two samples: H0: $\sigma_x = \sigma_y$ vs H0: $\sigma_x \neq \sigma_y$

- In R you use the function `var.test()`

```
> x = rnorm(56, 5, 2)
> y = rnorm(74, 4.5, 1.5)
> var.test(x, y)

        F test to compare two variances

data:  x and y
F = 1.8361, num df = 55, denom df = 73, p-value = 0.01532
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.124134 3.052659
sample estimates:
ratio of variances
          1.836106
```

Nonparametric tests

- Even though the previous tests assume normality, they are typically quite robust to the lack of it. Nonetheless, if the data is heavily skewed they might produce misleading results.
- Nonparametric tests avoid assumptions about the distribution of the data and are more generally applicable. The tradeofff is that they tend to have a lower power (i.e., they are more likely to miss a small signal in the data)

Nonparametric tests

- One of the most popular classes of nonparametric tests is the Wilcoxon/Mann-Whitney test.
- These procedures use the ranks of the observations to construct the test statistic.
  - The smallest observation has a rank of 1, the second smallest has a rank of 2, etc.
  - If two observations have the same value, their rank is the average of the ranks they would have taken.

Nonparametric tests

- The Wilcoxon test is analogous to the one sample t test:

  $$H0: \ \mu = \mu_0 \quad \text{vs} \quad H1: \ \mu \neq \mu_0$$

  (or $\mu < \mu_0$ or $\mu > \mu_0$) where μ is the median of the distribution of the data.

- The test statistic is given by the sum of the (signed) ranks of the observations below μ_0.

Nonparametric tests

- The Mann-Whitney test is analogous to a two-sample t test

  $$H0: \mu_1 = \mu_2 \quad \text{vs} \quad H1: \mu_1 \neq \mu_2$$

  (or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$) where μ_1 and μ_2 are the medians of both distributions.

- Again, the test is based on the the differences of ranks.

Nonparametric tests

- Both nonparametric tests are implemented in R through the `wilcox.test()` function:

```
> x = rnorm(56, 5, 2)
> y = rnorm(74, 4.2, 1.5)
> wilcox.test(x, y)

 Wilcoxon rank sum test with continuity correction

data:  x and y
W = 2457, p-value = 0.07064
alternative hypothesis: true location shift is not equal to 0

> t.test(x, y)

 Welch Two Sample t-test

data:  x and y
t = 1.793, df = 106, p-value = 0.07582
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05107093  1.01728721
sample estimates:
mean of x mean of y
 4.772491  4.289383
```

Categorical data and contingency tables

- A contingency table displays the joint frequency distribution of two or more categorical variables.
  - Think about it as an array with K dimensions, one per categorical variable. The size of each dimension corresponds to the number of categories for that variable.
- Categorical variables in R should be encoded as factors rather than numbers or characters.
- Count table can be generated using the `table()` or the `xtabs()` command.

Categorical data and contingency tables

- Make sure that you treat categorical variables as factors:

```
> arth = read.table("arthritis.txt", header=T)
> arth
   ID Treatment    Sex Age Improved
1  57   Treated   Male  27     Some
2  46   Treated   Male  29     None
3  77   Treated   Male  30     None
4  17   Treated   Male  32   Marked
> is.factor(arth[,2])
[1] TRUE
> summary(arth[,2:4])
  Treatment       Sex           Age
 Placebo:43   Female:59   Min.   :23.00
 Treated:41   Male  :25   1st Qu.:46.00
                          Median :57.00
                          Mean   :53.36
                          3rd Qu.:63.00
                          Max.   :74.00
```

Categorical data and contingency tables

- By having categorical variables defined as factors you can easily construct tables:

```
> arth = read.table("arthritis.txt", header=T)
> table(arth[,2], arth[,3])

          Female Male
  Placebo     32   11
  Treated     27   14
> table(arth[,2], arth[,3], arth[,5])

, ,  = Marked

          Female Male
  Placebo      6    1
  Treated     16    5

, ,  = None

          Female Male
  Placebo     19   10
  Treated      6    7

, ,  = Some

          Female Male
  Placebo      7    0
  Treated      5    2
```

Categorical data and contingency tables

- There is a version of the function that works with formula objects:

```
> arth = read.table("arthritis.txt",
header=T)
> attach(arth)
> xtabs(~Treatment+Sex)
          Sex
Treatment Female Male
  Placebo     32   11
  Treated     27   14
```

- xtabs can produce sparse 2D tables (useful for large datasets!)

Categorical data and contingency tables

- Let's focus on two-way contingency tables (i.e., those that involve the joint distribution of two variables)
- A common question of interest is whether these two variables are independent, i.e., whether $P(X = a, Y = b) = P(X = a)P(Y = b)$.
- Example: Is smoking more than two cigarettes a day associated with the development of cancer?

Categorical data and contingency tables

- Your data would look something like this

		Cancer?	
		Yes	**No**
Smoke?	**Yes**	n_{11}	n_{12}
	No	n_{21}	n_{22}

- How would you expect the table to look like if the two variables where independent?

Categorical data and contingency tables

- First compute the margins (row and column sums to the observed values)

		Cancer?		
		Yes	**No**	
Smoke?	**Yes**	n_{11}	n_{12}	$n_{1\bullet} = n_{11} + n_{12}$
	No	n_{21}	n_{22}	$n_{2\bullet} = n_{21} + n_{22}$
		$n_{\bullet 1} = n_{11} + n_{21}$	$n_{\bullet 2} = n_{12} + n_{22}$	$n_{\bullet\bullet} = n_{11} + n_{12} + n_{21} + n_{22}$

Categorical data and contingency tables

- Create a table that has the same margins but where entries are independent

		Cancer?		
		Yes	**No**	
Smoke?	**Yes**	$n_{1\bullet}n_{\bullet 1}/n_{\bullet\bullet}$	$n_{1\bullet}n_{\bullet 2}/n_{\bullet\bullet}$	$n_{1\bullet} = n_{11} + n_{12}$
	No	$n_{2\bullet}n_{\bullet 1}/n_{\bullet\bullet}$	$n_{2\bullet}n_{\bullet 2}/n_{\bullet\bullet}$	$n_{2\bullet} = n_{21} + n_{22}$
		$n_{\bullet 1} = n_{11} + n_{21}$	$n_{\bullet 2} = n_{12} + n_{22}$	$n_{\bullet\bullet} = n_{11} + n_{12} + n_{21} + n_{22}$

Check that the sums of rows and columns are the same as in the original table!

Categorical data and contingency tables

- The test statistic is the difference between the observed and the expected values

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{i,j}-\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}} \sim \chi_1^2$$

Valid as long as $\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}$ is not too small

- This is the Pearson's χ^2 test.
- The same idea can be used for two-way contingency tables with more than two levels each. The number of degrees of freedom for the χ^2 distribution is (r-1)(c-1) (c = number of columns, r = number of rows).

Categorical data and contingency tables

- You can run this test using the function `chisq.test()`:

```
> arth = read.table("arthritis.txt", header=T)
> arth.tab = table(arth[,2], arth[,5])
> arth.tab

          Marked None Some
  Placebo      7   29    7
  Treated     21   13    7

> chisq.test(arth.tab)

 Pearson's Chi-squared test

data:  arth.tab
X-squared = 13.055, df = 2, p-value = 0.001463
```

- Testing for higher order tables requires the use of GLMs.