Introduction to Statistics with R

Presenter: Steve Baskauf
steve.baskauf@vanderbilt.edu

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Factors and the t-test of means

Presenter: Steve Baskauf
steve.baskauf@vanderbilt.edu

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS
Jean & Alexander Heard LIBRARIES

CodeGraf landing page

- vanderbi.lt/codegraf

Factors and experimental design

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Factors

- A **factor** is a data structure for categorizing discontinuous data.
- Its origin comes from **experimental design** terminology.
- In an experiment, each **category** into which an experimental trial can fall is called a **level**.
- Factors are sometimes called **grouping variables** because they are used to group observations.
- Factors may be required for some statistical tests and visualizations.

Factor example: science fair

water factor	height (cm)
wet	25
wet	21
dry	14
wet	13
dry	10
wet	18

- The water factor has two levels: wet and dry
- The height observations can be grouped by whether the experimental treatment was wet or dry

Factor example: creating factor values

- Create a vector of character strings and a vector of number values:

```
water_conditions <- c("wet", "wet", "dry",
"wet", "dry", "wet")

height <- c(25, 21, 14, 13, 10, 18)
```

- Convert the strings into a factor

```
water_factor <- factor(water_conditions)
```

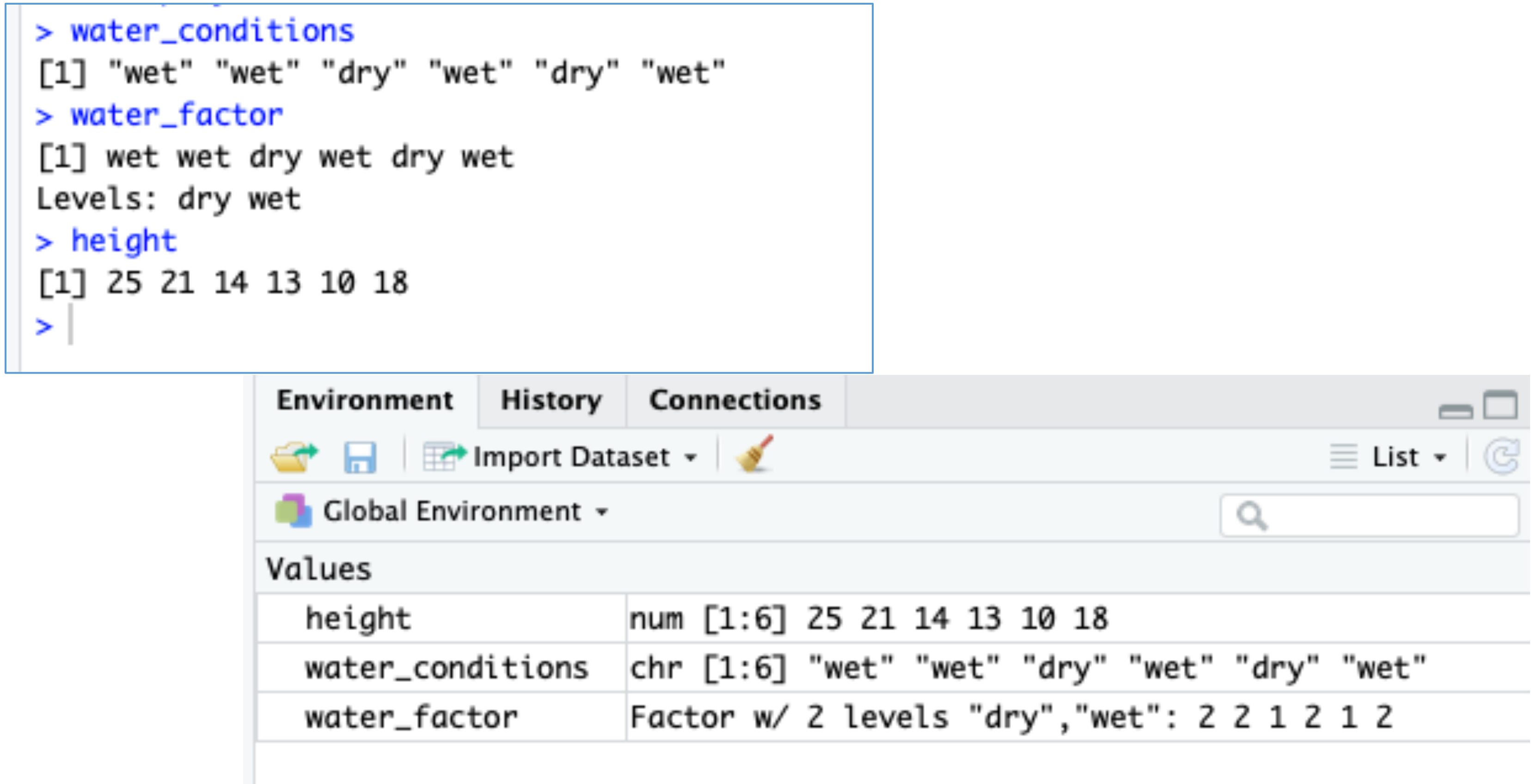
- Display the values of each data structure

```
water_conditions

water_factor

height
```

How to tell that a data structure is a factor



- The main clue is that the **values of the levels** are listed.
- Notice that the levels are actually stored as numbers. The factor level strings are just labels for the numbers.

Data frames and factors

- **character strings** are automatically turned into **factors** when data frames are built from individual vectors.

```
group <- c("reptile", "arachnid", "annelid", "insect")
# character strings
number_legs <- c(4,8,0,6) # numbers
organism_info <- data.frame(group, number_legs)
```

- This can be good or bad depending on how you want to use the data.

Data frames and factors

- Recall that **character strings** read from CSV files using `read.csv()` are automatically turned into **factors**
- **numbers** imported from CSV files are imported as **number vectors**
- This automatic behavior takes place because of the historical orientation of R towards statistics.
- Use `as.character()` to convert from factors to character strings, or read in as a tibble using `read_csv()`

t-test of means

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

t-test of means characteristics

- The **independent variable** is **discontinuous** (factor)
- The **dependent variable** is **continuous** (numeric)
- The **factor** (a.k.a. grouping variable) has only **2 levels**.
- We want to know if the two levels of the independent variable have significantly different means for the continuous variable.

t-test of means applications

- This test is great for a controlled manipulative experiment or "randomized controlled trial")

water factor	height (cm)
wet	25
wet	21
dry	14
wet	13
dry	10
wet	18

- Is the mean height for **`wet`** (=level of factor) different from the mean for **`dry`**?

t-test of means in R

- Need two vectors (or columns from data frame).
- One (independent variable) must be a factor, the other (dependent variable) must be numeric.
- Must be organized as "tidy data" (category data in a single column)
- Format:

```
t.test(dep_vec ~ ind_vec, var.equal=TRUE)
```

or

```
t.test(dep_col ~ ind_col, data=data_frame, var.equal=TRUE)
```

Review of p-value

DISC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Result of Two Sample t-test on heights data

```
data:  height by grouping
t = 2.7654, df = 12, p-value = 0.01711
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
  1.869768 15.758803
sample estimates:
  mean in group men mean in group women
           179.8714            171.0571
```

Sampling from a population

- In the heights data, we have 7 heights for each sex.
- The data represent a **sample** of heights from the populations of men and women (all possible men and all possible women).

```
mean in group men mean in group women
         179.8714            171.0571
```

- Possibilities:
  - men and women have the same average height, but by chance our sampling was unrepresentative (null hypothesis)
  - the sampling was representative of the populations and heights of men and women are different (alternative hypothesis)

What is a p-value (P) ?

- What is P?
  - P is the probability that we would see results like this if nothing interesting were going on (variation is random).
  - P = 0.6 (could be like this 60% of the time if random; likely to be random)
  - P = 0.001 (could be like this 0.1% of the time if random; not likely to be random)
- If it's really unlikely that our results would occur when only random things are happening, we think something interesting is going on.
- P is an assessment of the null hypothesis (nothing interesting)
- When P is low, we reject the idea that nothing interesting is going on (the null hypothesis)

Why do we like it when P<0.05 ?

- Hypotheses:
  - things are different (alternate hypothesis)
  - things are the same (null hypothesis)
- Strategy:
  - show that the null hypothesis is wrong
- If $P < 0.05$, then we assume the null hypothesis is wrong because it's so unlikely.
- If $P > 0.05$, then either the null hypothesis is correct or our experiment STINKS !
- We probably know what's going on if $P < 0.05$ but not if $P > 0.05$

Statistical power

- Power is the ability to show that different things are different ($P < 0.05$)
- We get more statistical power if there's less variation in the data or a larger sample size.
- We may be able to control variation by experimental conditions
- We should be able to increase the sample size (if we have time and money).
- If not different, increasing power won't reduce P.
- If different, increasing power will make P get smaller.

Power tradeoff

- Too little statistical power:
  - can't show that different things are different
  - Unable to get $P > 0.05$ when there are differences.

- It seems like more statistical power would always be a good thing, but…
- Too much statistical power
  - tiny unimportant things are shown to be different
  - $P < 0.05$ for factors with a very small effect.

Assumptions of t-test of means

DiSC DIGITAL SCHOLARSHIP AND COMMUNICATIONS

Jean & Alexander Heard LIBRARIES

Testing assumption of t-test of means

1. Independence of the two samples (not testable by stats).
2. Each group normally distributed.
3. Variances of the two groups are the same.

(Typical requirements for **parametric tests**.)

Testing assumption of normality

- Graphical examination:
  - histogram
  - normal quantile plot
- Shapiro-Wilkes test

Testing assumption of equal variances

- Bartlett's test
  - built-in function
  - valid if data are normally distributed