Lecture 14

Ciprian M Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

November 2, 2021

Table of contents

F-test description

1. T-test is used to compare the means of two groups
2. Sometimes we want to compare the means of multiple groups
3. Consider K groups with independent observations. The observations in the kth group are

$$X_{11}, \ldots, X_{1k} \sim N(\mu_k, \sigma^2)$$

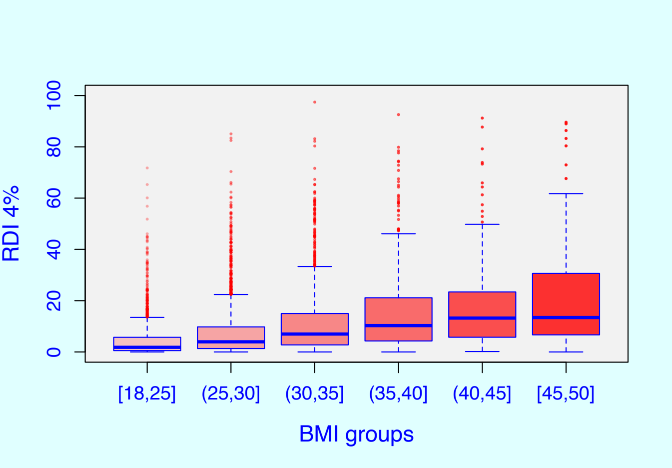
4. We want to test the global hypothesis

$$H_0 : \mu_1 = \ldots = \mu_K = \mu$$

against the alternative H_A that at least two means are equal

Example: SHHS

- Distribution of RDI 4% in six BMI groups.



Intuition behind F statistic

1. Let $\overline{X}_{.k}$ be the mean of the kth group
2. Let $\overline{X} = \sum_k n_k \overline{X}_{.k}/n$ the mean of all observations
3. Under the null, one expects that $\overline{X}_{.k}$ are close together and to $\overline{X}$
4. The F statistic

$$Y = \frac{\sum_{k=1}^{K} n_k(\overline{X}_{.k} - \overline{X})^2/(K-1)}{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_{ik} - \overline{X}_{.k})^2/(n-K)}$$

5. Numerator: a measure of how far the group means $\overline{X}_{.k}$ are from the population mean $\overline{X}$ (between group variance)
6. Denominator: a measure of how far the individual observations are from their respective group means (within group variance)

F statistic: numerator

1. Under the null $\overline{X}_{.k} \sim N(\mu, \sigma^2/n_k)$
2. $\frac{n_k(\overline{X}_{.k}-\mu)^2}{\sigma^2} \sim \chi^2_1$
3. $n_k(\overline{X}_{.k} - \overline{X})$ are independent
4. $\sum_{k=1}^K \frac{n_k(\overline{X}_{.k}-\overline{X})^2}{\sigma^2} \sim \chi^2_{K-1}$
5. One degree of freedom is "lost" from replacing μ by $\overline{X}$
6. The numerator is $Y_{K-1}/(K-1)$, where $Y_{K-1} \sim \chi^2_{K-1}$

F statistic: denominator

1. $\sum_{i=1}^{n_k} \frac{(X_{ik}-\overline{X}_{.k})^2}{\sigma^2} \sim \chi^2_{n_k-1}$
2. $\sum_{i=1}^{n_k} \frac{(X_{ik}-\overline{X}_{.k})^2}{\sigma^2}$ are independent
3. Hence $\sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{(X_{ik}-\overline{X}_{.k})^2}{\sigma^2} \sim \chi^2_{n-K}$
4. The denominator is $Y_{n-K}/(n-K)$, where $Y_{n-K} \sim \chi^2_{n-K}$
5. The numerator and denominator are independent

Back to the F statistic

1. Because σ^2 cancels out, under the null hypothesis

$$Y = \frac{Y_{K-1}/(K-1)}{Y_{n-K}/(n-K)}$$

2. $Y_{K-1} \sim \chi^2_{K-1}$ and $Y_{n-K} \sim \chi^2_{n-K}$ are independent
3. The distribution of this variable is called the F-distribution with (K-1,n-K) degrees of freedom
4. Reject the null hypothesis if the F statistic is too large

```
qf(0.95,5,5755)
[1] 2.215653
```

Example: F test

1. We apply the F test for testing the null hypothesis that the mean RDI 4% are the same in the six BMI groups

```
one.way<-aov(rdi4p~bmi_cut)
> summary(one.way)
              Df Sum Sq Mean Sq F value
bmi_cut        5  85066   17013   121.2
Residuals   5755 807641     140
            Pr(>F)
bmi_cut     <2e-16 ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
43 observations deleted due to missingness
```

F-test context

- This is a global test
- Rejecting the null does not provide information about which two means are not equal
- In the case of two groups the F test is the square of the t-test
- The F test in this context is the one way ANOVA (analysis of variance)
- You will see it again in regression when comparing two nested models

One and two way ANOVA

```
one.way<-aov(rdi4p~bmi_cut)

two.way<-aov(rdi4p~bmi_cut+gender)
> summary(two.way)
              Df Sum Sq Mean Sq F value
bmi_cut        5  85066   17013   127.8
gender         1  41448   41448   311.3
Residuals   5754 766193     133
            Pr(>F)
bmi_cut     <2e-16 ***
gender      <2e-16 ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
43 observations deleted due to missingness
```

Connection to regression

- One way and two way ANOVA are linear regressions with a continuous outcome (RDI 4%)
- One way ANOVA has one categorical regressor (BMI categories)
- Two way ANOVA has two categorical regressors (BMI categories and gender)

Reasons for data transformations

- Distributions contain extreme outliers, skewness
- Concerns that statistical properties may not hold
- Harmonization across studies
- Concerns that errors may not be additive

Main types of transformations

- Z-scoring (standardization)

$$Z = \frac{X - \mathrm{E}(X)}{\mathrm{SD}(X)}$$

- The Box-Cox family of transformations ($\lambda \geq 0$)

$$Y_\lambda = \frac{X^\lambda - 1}{\lambda}$$

- Log transformation ($\lambda \downarrow 0$): $Y = \log(X)$
- Square root transformation ($\lambda = 1/2$): $Y = \sqrt{X}$

Main alternatives

- Sensitivity analyses (remove top outliers and rerun analyses)
- Nonparametric (quantile analyses)

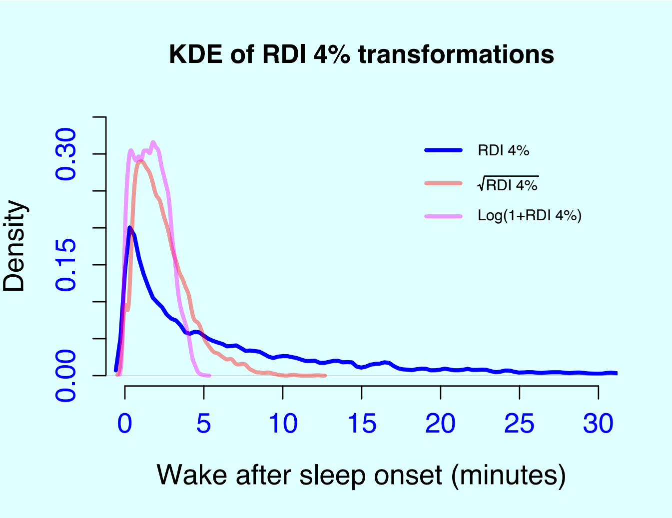
Main drawbacks of transformations

- Data are no longer on the original scale
- Data interpretation is changed
- If $Y = h(X)$ is a generic transformation of the rv X

$$E\{h(X)\} \neq h\{E(X)\}$$

- Thus, transforming the data and taking the mean and then transforming back does not give you the original mean

$$h^{-1}[E\{h(X)\}] \neq E(X)$$


KDE of RDI 4% transformations
Density
0.00
0.15
0.30
RDI 4%
√RDI 4%
Log(1+RDI 4%)
0
5
10
15
20
25
30
Wake after sleep onset (minutes)

The log-normal distribution

- A random variable is **log-normally** distributed *if its log is a normally distributed random variable*
- "I am log-normal" means "take logs of me and then I'll then be normal"
- Note log-normal random variables are not logs of normal random variables!!!!!! (You can't even take the log of a normal random variable)
- Formally, X is lognormal(μ, σ^2) if $\log(X) \sim \mathrm{N}(\mu, \sigma^2)$
- If $Y \sim \mathrm{N}(\mu, \sigma^2)$ then $X = e^Y$ is log-normal

The log-normal distribution

- The log-normal density is

$$\frac{1}{\sqrt{2\pi}} \times \frac{\exp[-\{\log(x) - \mu\}^2/(2\sigma^2)]}{x} \quad \text{for} \quad 0 \leq x \leq \infty$$

- Its mean is $e^{\mu + (\sigma^2/2)}$ and variance is $e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$
- Its median is e^{μ}

The log-normal distribution

- Notice that if we assume that $X_1, \ldots, X_n$ are log-normal(μ, σ^2) then $Y_1 = \log X_1, \ldots, Y_n = \log X_n$ are normally distributed with mean μ and variance σ^2
- Creating a Gosset's t confidence interval on using the Y_i is a confidence interval for μ the log of the median of the X_i
- Exponentiate the endpoints of the interval to obtain a confidence interval for e^μ, the median on the original scale
- Assuming log-normality, exponentiating t confidence intervals for the difference in two log means again estimates ratios of geometric means

Example: interpret these results

Gray matter volumes for 342 older subjects (over 60) and 287 younger subjects were compared.

- The mean log gray matter volumes was 6.35 $\log(\mathrm{cm}^3)$ (older) and 6.40 $\log(\mathrm{cm}^3)$ (younger). Exponentiating these numbers leads to 570.90 cm^3 and 599.40 cm^3
- The SDs were 0.11 $\log(\mathrm{cm}^3)$ and 0.11 $\log(\mathrm{cm}^3)$
- CIs
  - Younger: log scale - $[6.38, 6.41]$, exponentiated - $[592.03, 606.86]$
  - Older: log scale - $[6.34, 6.36]$, exponentiated - $[564.36, 577.50]$
- Two sample mean comparison
  - Log scale - $[0.03, 0.07]$
  - Exponentiated - $[1.03, 1.07]$