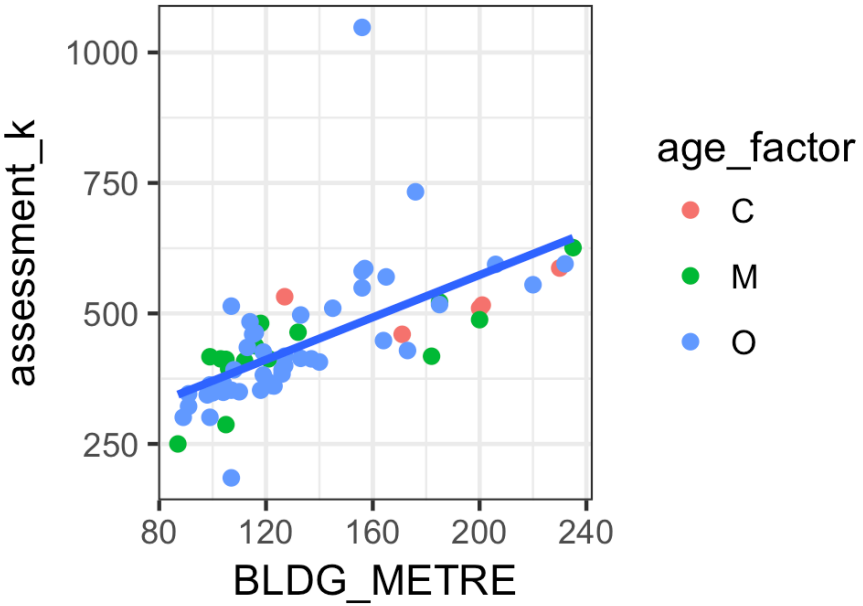
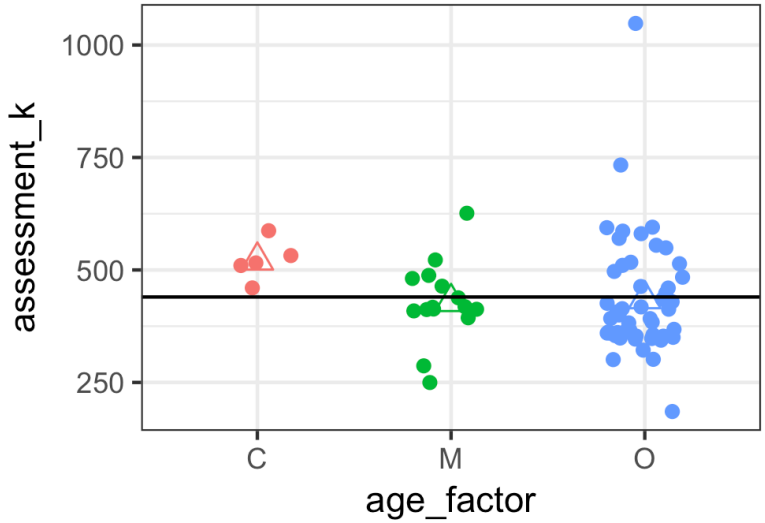
DSCI561: Regression I
Lecture 1: November 15, 2017

Gabriela Cohen Freue
Department of Statistics, UBC

A **regression** analysis is used to model the relationship between a *dependent* variable and one or more *independent* variables.

dependent variable: also called `response` (*y*)





independent variables: also called `explanatory variables`, `covariates`, `predictors` (*X*)

Outline of the course

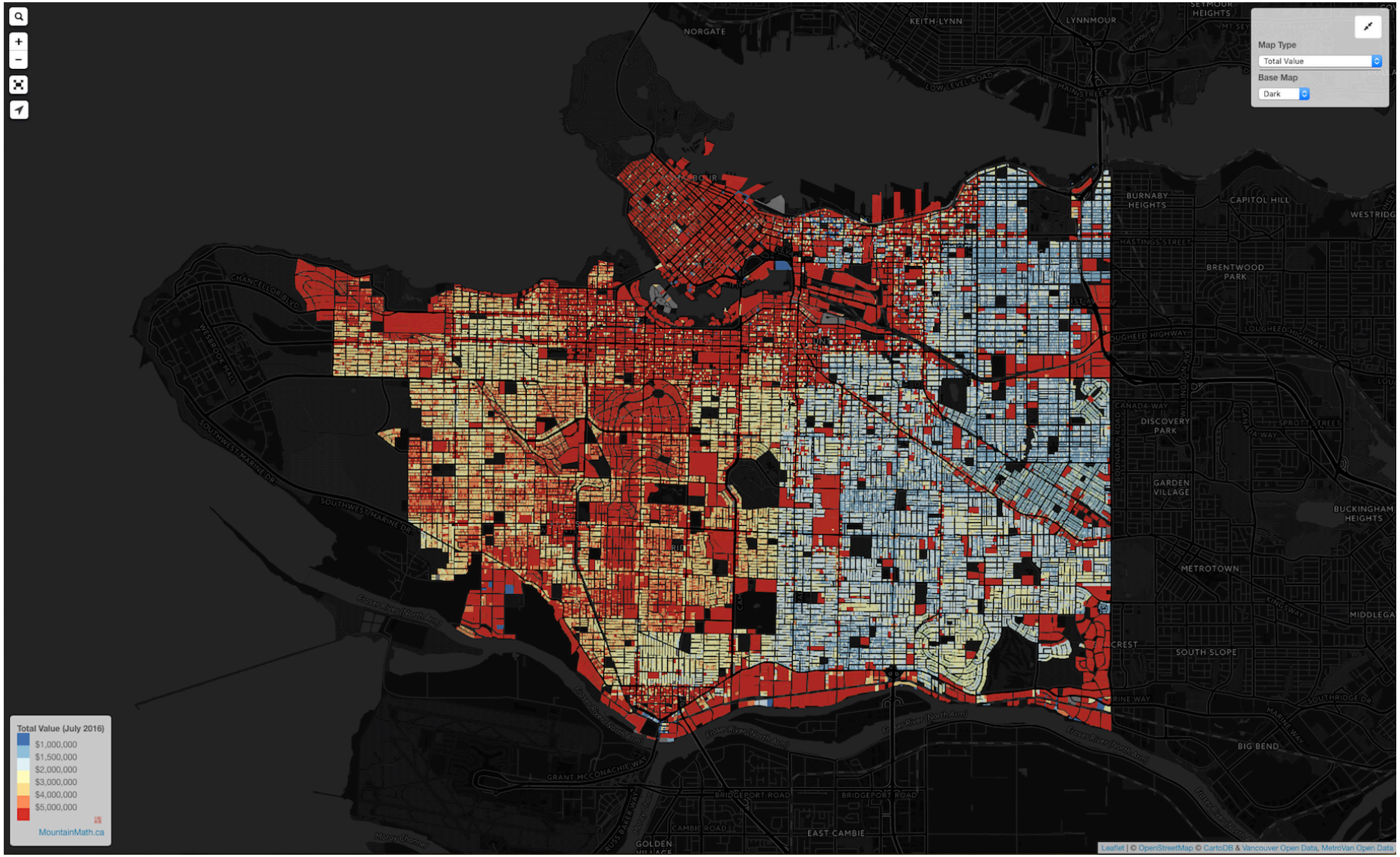
Note: this is only a preliminary outline. Expect some changes.

- **Lecture 1**
  - Review two-sample t-tests
  - Review ANOVA
  - Connection between two-samples t-test, ANOVA
- **Lecture 2**
  - Review some linear algebra operations
  - Connection between two-samples t-test, ANOVA and regression
- **Lecture 3**
  - Unleash the scope of regression analysis: many independent variables of different kind

Outline of the course (cont.)

- **Lecture 4**
  - Estimation in linear regression
  - Confidence and prediction intervals
- **Lectures 5 and 6**
  - Bootstrap
  - Permutation test
- **Lecture 7**
  - Residuals and goodness of fit
  - Multicolinearity
- **Lecture 8**
  - Diagnostics

Property Assessment Tax Data



Map: https://moutainmath.ca

Contemporary

Modern

Old

Does the age of the property affect its tax assessment?

DATA

- 2015 Tax Property Assessment Data from Strathcona County
- Raw data: OpenDataBC (https://www.opendatabc.ca/datasets/property-tax-assessment-strathcona-county)
- Filter residential properties

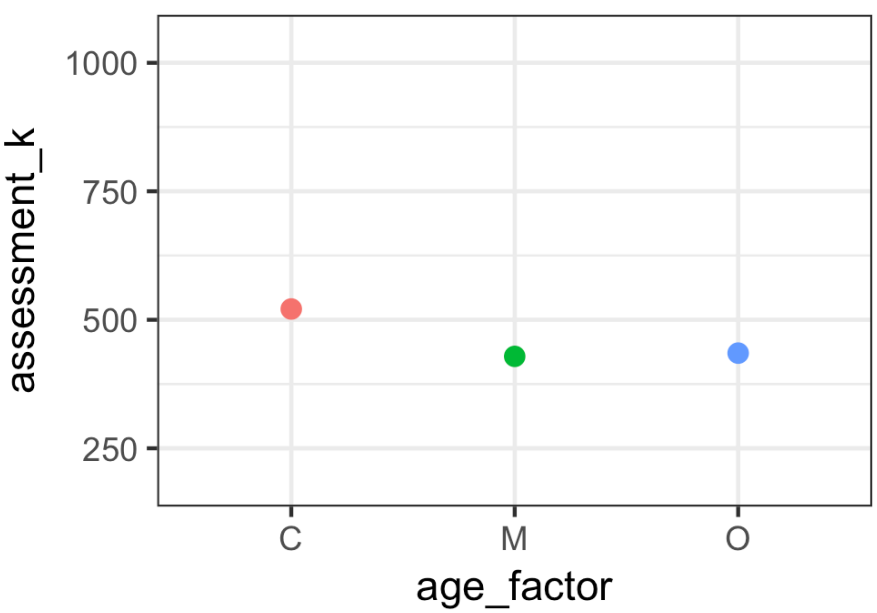
TAX_YEAR	YEAR_BUILT	BLDG_DESC	BLDG_METR	BLDG_FEET	GARAGE	FIREPLACE	BASEMENT	BSMTDEVL	ASSESSMENT
2015	1950	1 Storey & Basement	83	895	Y	Y	Y	Y	402000
2015	1983	1 Storey & Basement	77	831	N	Y	Y	N	292000
2015	1981	Split Entry	161	1731	N	Y	Y	Y	518000
2015	1967	1 Storey Basementless	56	600	Y	Y	N	N	197000
2015	1968	1 1/2 Sty. Slab on Grade	52	560	N	Y	N	N	181000
2015	...	...	...	...	...	...	...	...	...
2015	1962	1 1/2 Storey & Basement	95	1021	Y	Y	N	N	302000

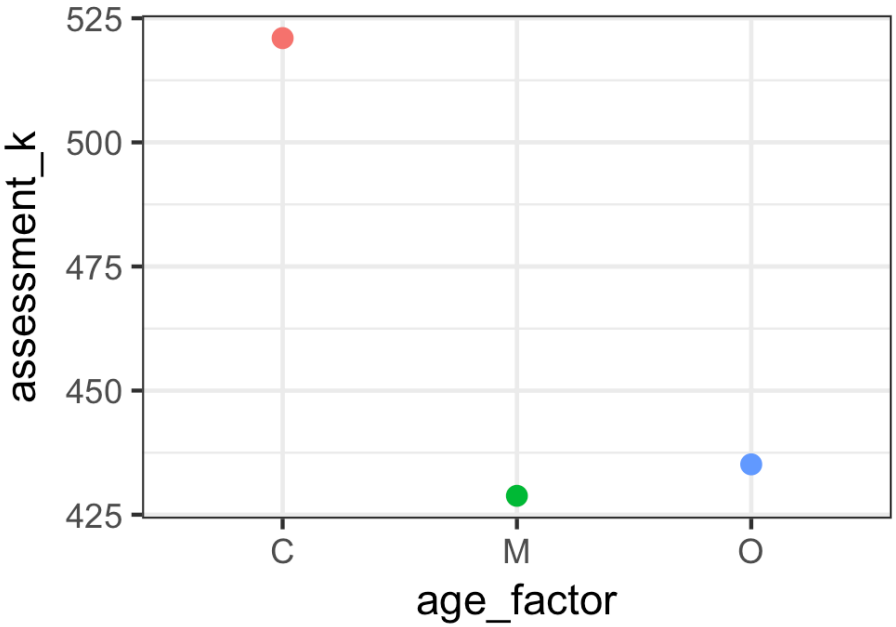
DATA (cont.)

```
> dat<-read.csv("dataPropTaxAssess.csv")
>
> # Add age of property
> dat$age<-2017-as.numeric(dat$YEAR_BUILT)
> dat$age_factor <- cut(dat$age,c(0,17,37,120),labels=c("C","M","O"))
> dat <- dat %>% mutate(assessment_k = ASSESSMENT / 1000)
> dat %>% select(age_factor) %>% summary
 age_factor
 C: 8785
 M: 8570
 O:10388
> dat %>% group_by(age_factor) %>% summarize(avg_assessment=mean(ASSESSMENT))
# A tibble: 3 x 2
  age_factor avg_assessment
      <fctr>          <dbl>
1          C       574978.9
2          M       515779.8
3          O       413212.9
> dat.small.C<-dat %>% filter(age_factor=="C")
> dat.small.M<-dat %>% filter(age_factor=="M")
> dat.small.O<-dat %>% filter(age_factor=="O")
>
> set.seed(123)
> dat.small<-rbind(dat.small.C[sample(1:nrow(dat.small.C),5),],
+                  dat.small.M[sample(1:nrow(dat.small.M),15),],
+                  dat.small.O[sample(1:nrow(dat.small.O),50),])
```

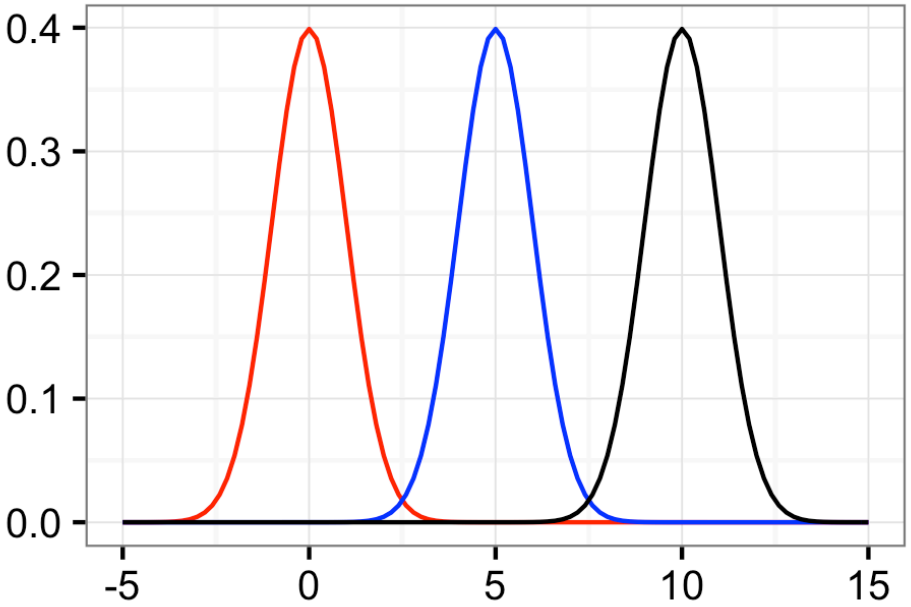
small sample
for illustration

Average tax value in my small sample by property age



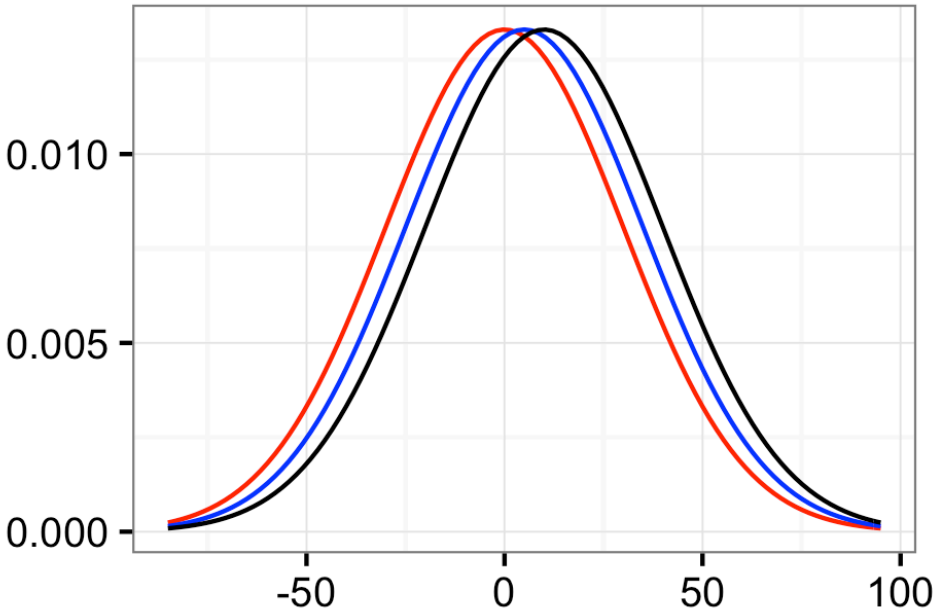


Does the age of the property affect its tax assessment?

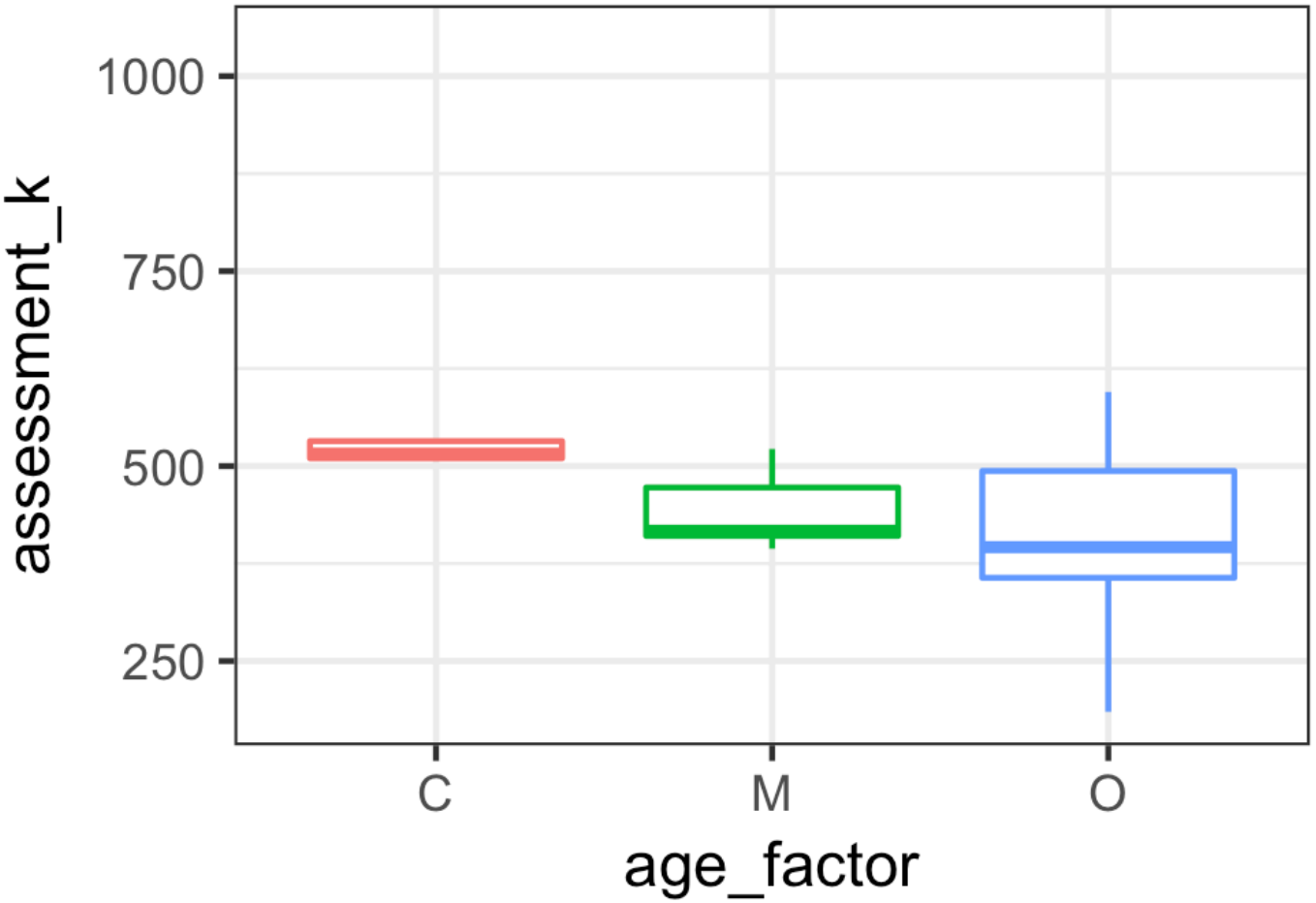


same centers

different variation

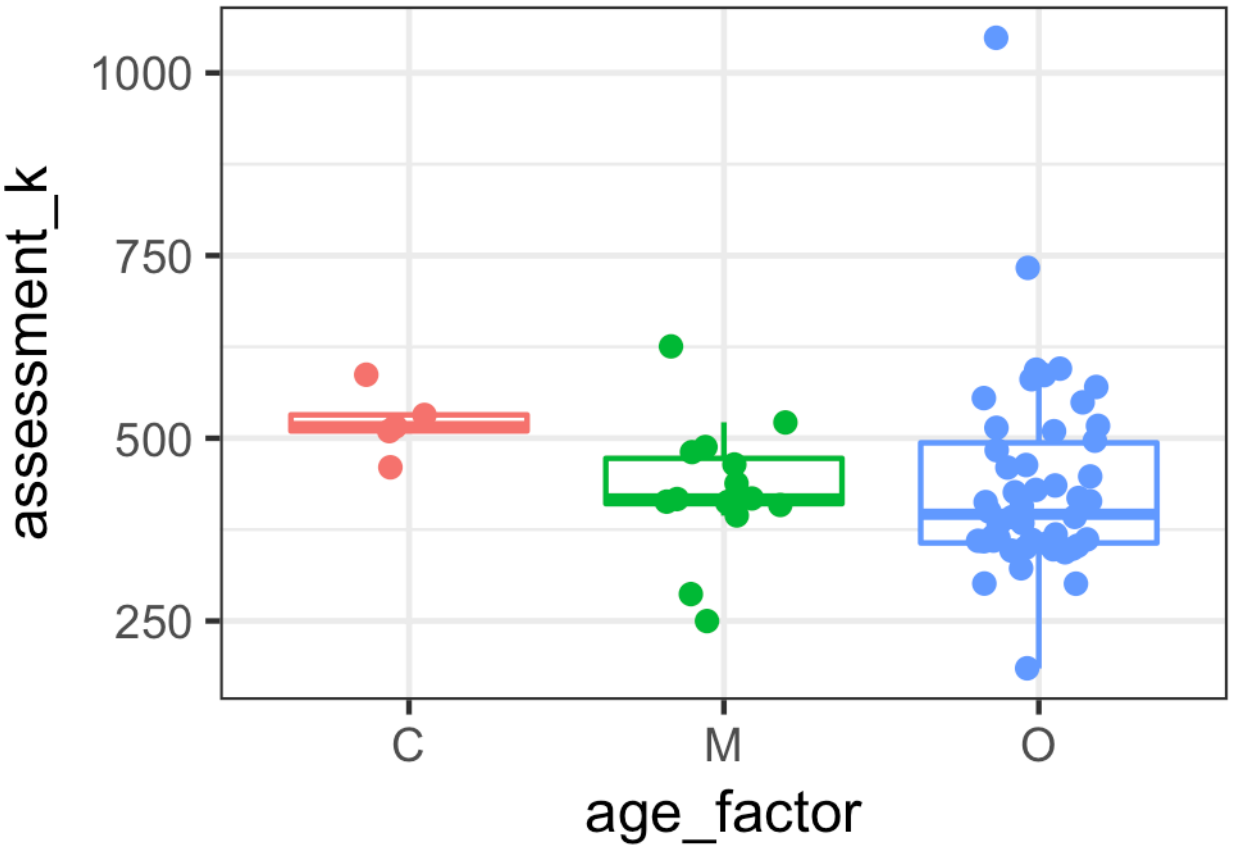


Boxplots of tax value in my small sample by property age



Does the age affect tax value?

Data points: tax value in my small sample by property age



Does the age affect tax value?

Notation

Random variables

Y_i : tax assessment value (`assessment`) for properties built in period i

$Y_{i1},\ Y_{i2}, \ldots, Y_{in_i}$: a random sample of size n_i of properties built in period i

$\bar{Y}_i = \dfrac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$: sample mean of tax values of properties built in period i

Population parameters

$\mu_i = E[Y_i]$: the (population) expected tax value for a property built in period i

Glance at the data

YEAR_BUILT	age_factor	assessment_k
2013	C	510
2003	C	516
2002	C	460
2002	C	532
2005	C	587
1995	M	481
1989	M	409
1991	M	522
1998	M	413
1989	M	413
1988	M	394
1990	M	418
1998	M	464
1990	M	412
1989	M	488
1990	M	626
1984	M	417
1989	M	250
1997	M	438
1980	M	287
1970	O	360
1974	O	426
1978	O	413
1976	O	460

l to 25 of 70 entries

$$Y_C;\ Y_{C1}, \ldots, Y_{C5},\ n_C = 5$$

$$Y_M;\ Y_{M1}, \ldots, Y_{M15},\ n_M = 15$$

$$Y_O;\ Y_{O1}, \ldots, Y_{O50},\ n_O = 50$$

Statistical Inference

- Which factors affect the expected value of the tax assessments in the "*population*"?
- **Hypothesis** (H_0): Age does not affect tax assessment

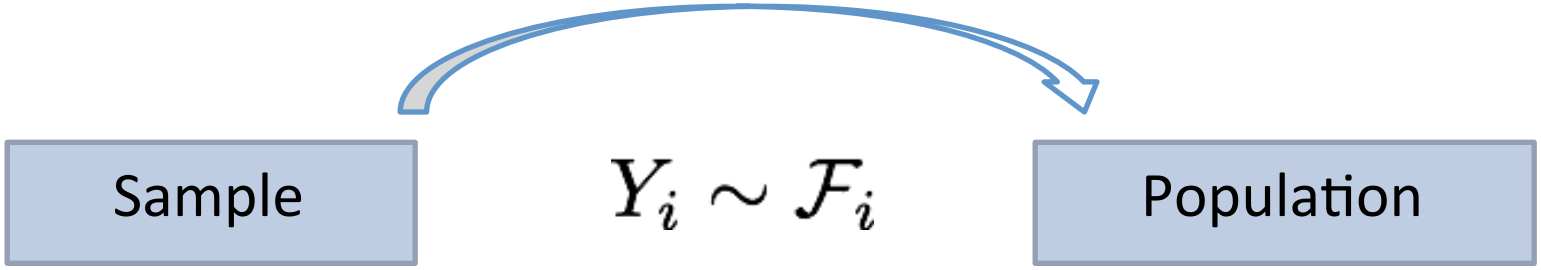
$$H_0 : \mu_C = \mu_M = \mu_O$$

- To test this hypothesis, we can *sample* properties from the 3 age periods

Statistical Inference

- Define 3 random variables: $Y_C;\ Y_M;\ Y_O$

  and collect a sample of tax values for contemporary (C), modern (M) and old (O) houses.



$$\hat{\mu}_C = \bar{Y}_C$$

$$\mu_C = E[Y_C]$$

$$\frac{\bar{Y}_M - \bar{Y}_C}{S_p\sqrt{\frac{1}{n_M} + \frac{1}{n_C}}};$$

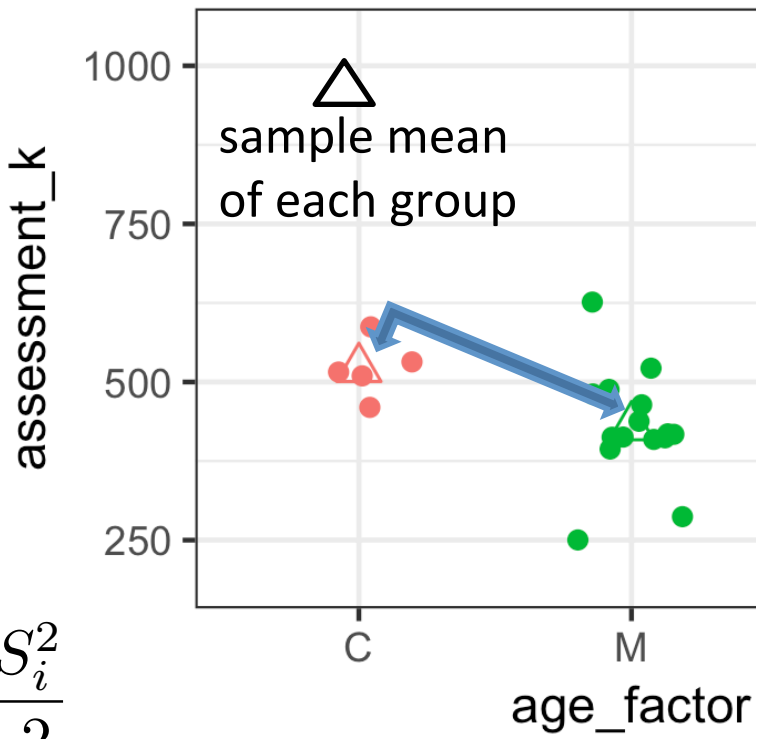
$$H_0 : \mu_C = \mu_M$$

Two sample t-test: compare 2 age periods

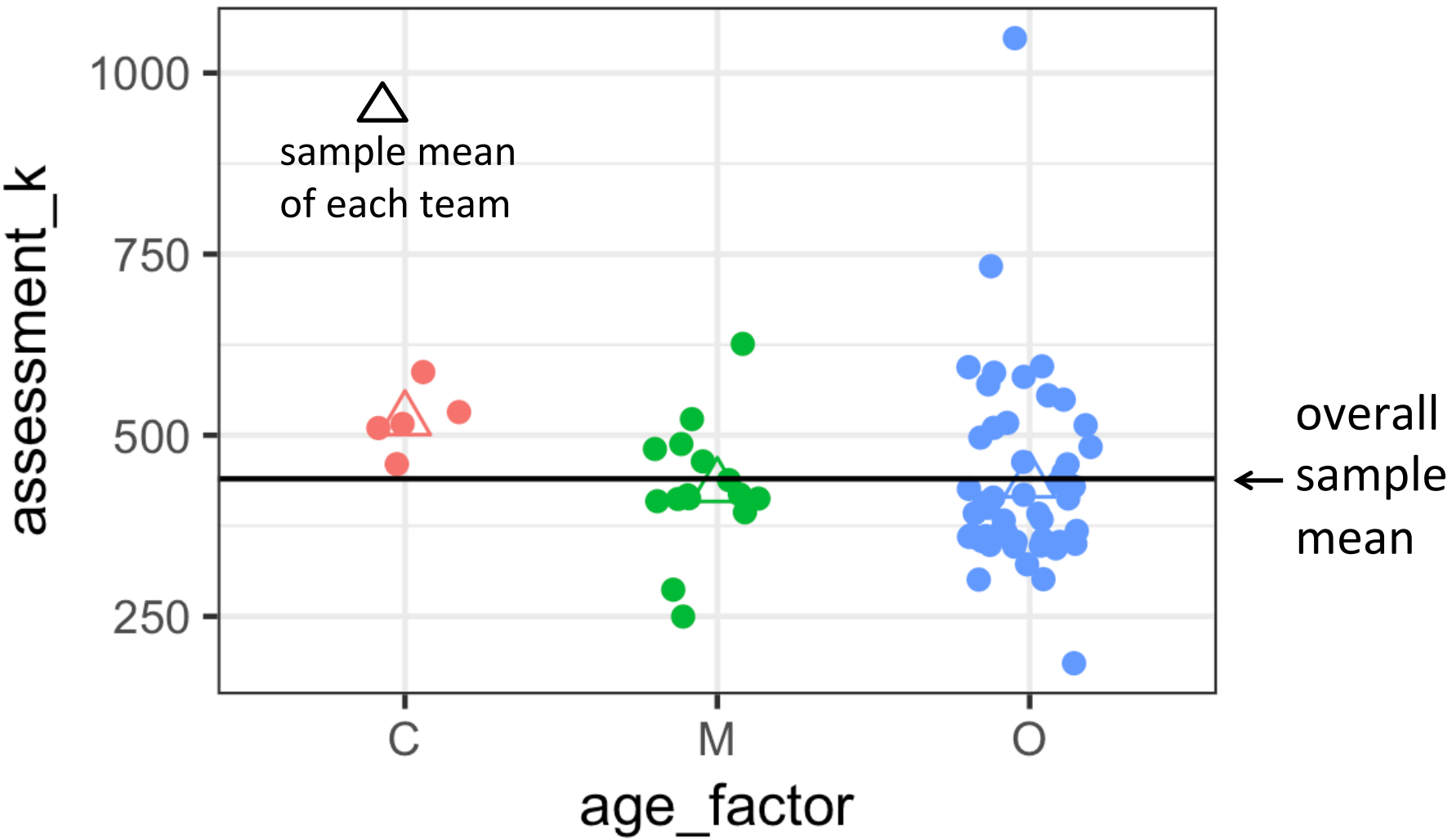
$$t = \frac{\bar{Y}_M - \bar{Y}_C}{\sqrt{\frac{S_M^2}{n_M} + \frac{S_C^2}{n_C}}}$$

or, assuming equal population variances:

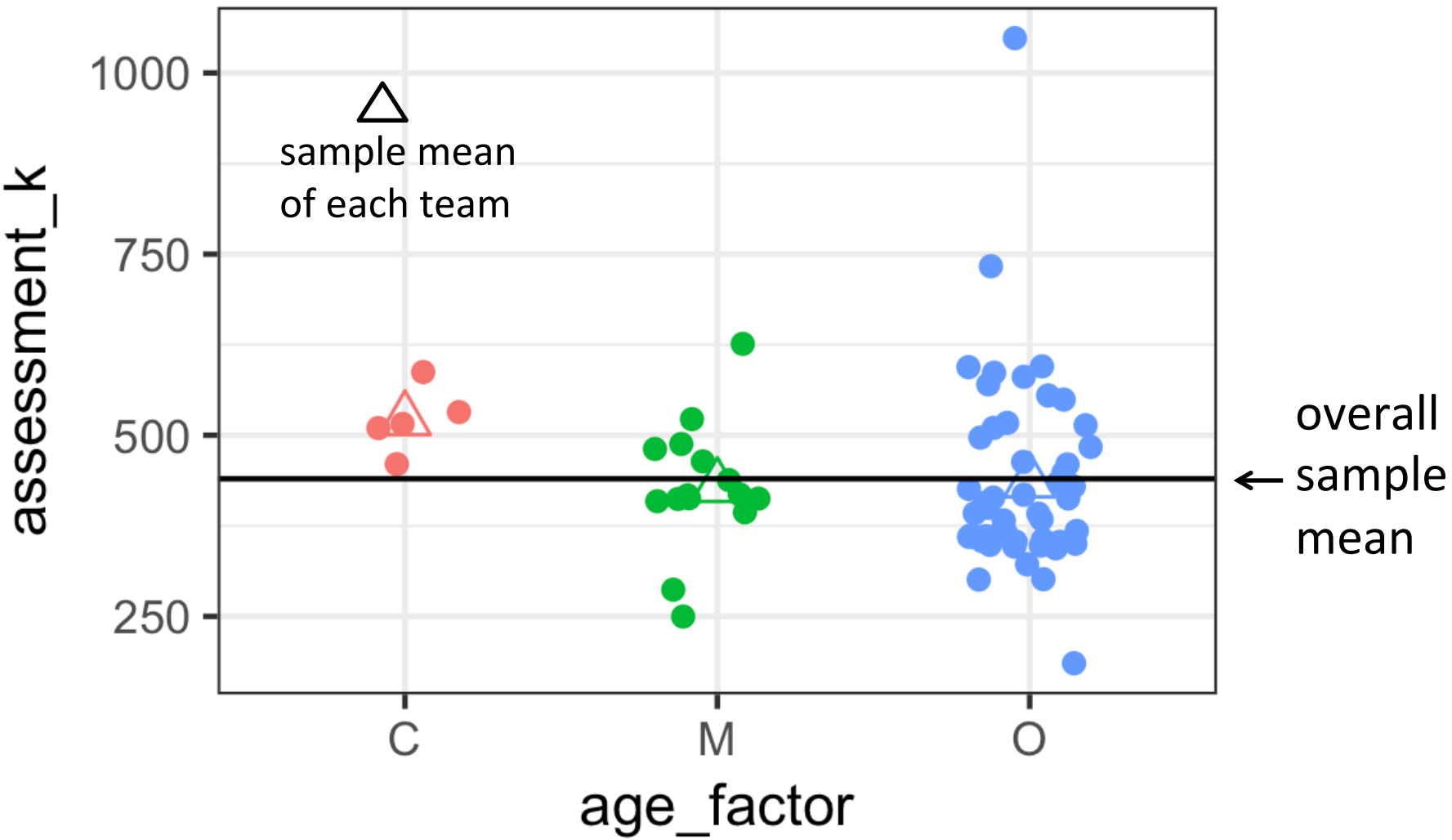
$$t = \frac{\bar{Y}_M - \bar{Y}_C}{S_p\sqrt{\frac{1}{n_M} + \frac{1}{n_C}}};\ S_p^2 = \frac{\sum_i^2 (n_i - 1)S_i^2}{n_M + n_C - 2}$$



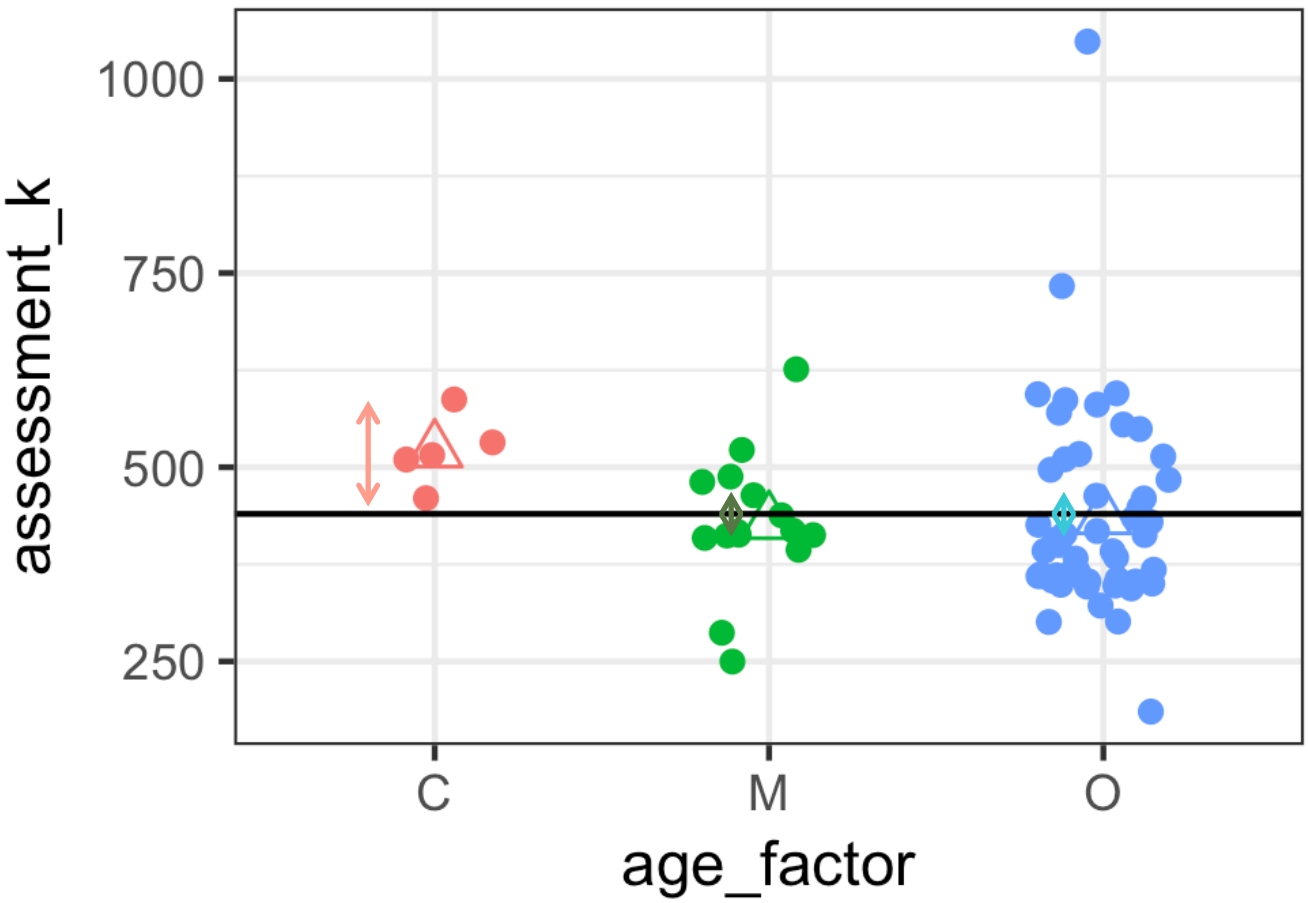
To test difference in ***population*** means we study the difference of the ***sample*** means relative to the SD of this difference



ANOVA: can compare more than 2 groups



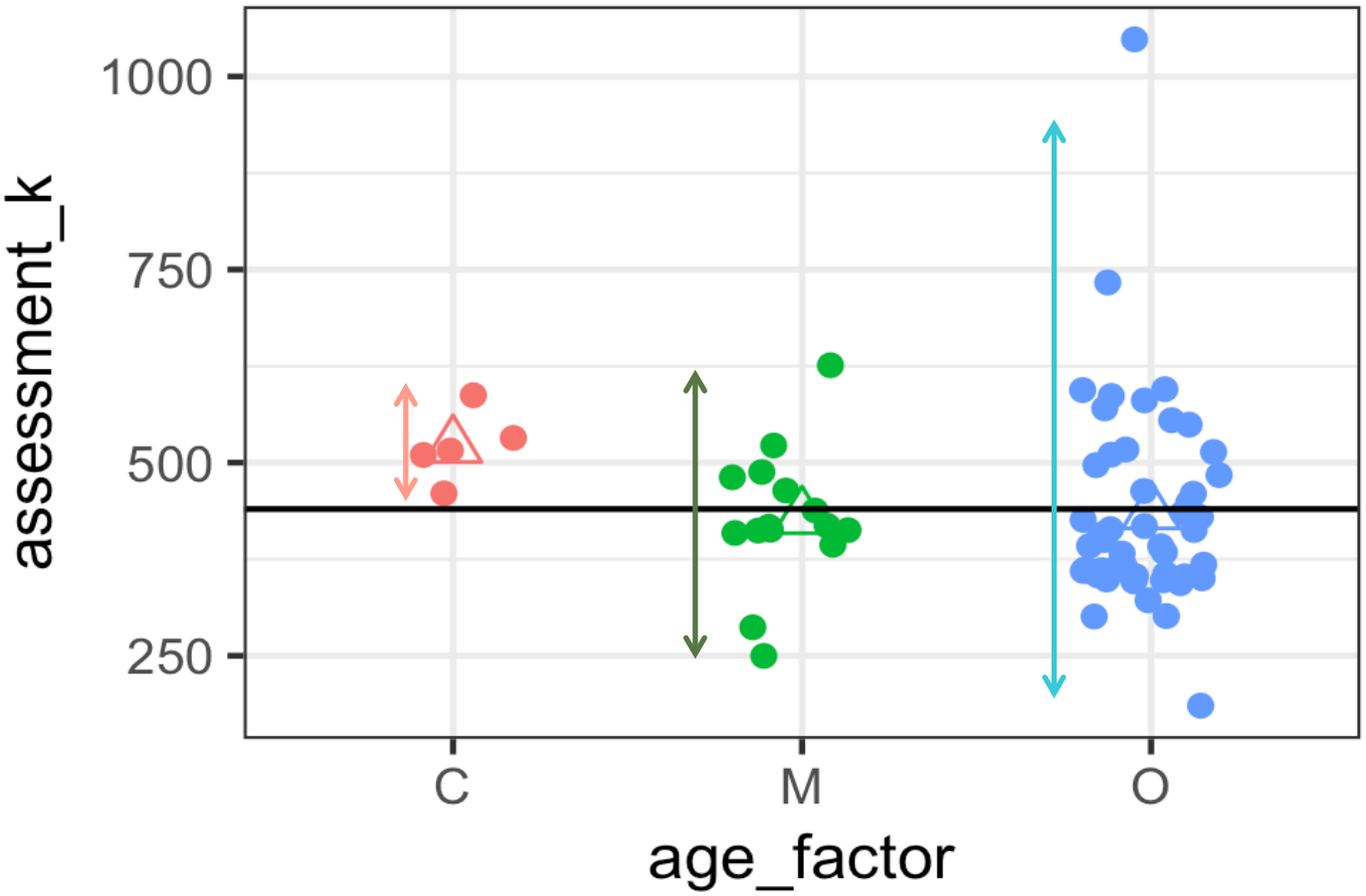
To test the difference among population means … why do we need to **analyze the variances** (ANOVA)??



To test the difference among population means we study the **between group variation …**

$$MSB = \frac{\sum_i^K n_i(\bar{Y}_i - \bar{Y})^2}{K-1}$$

NOTE: MSB: mean squares between, since it is the sum of squares between groups divided by its degrees of freedom. Note that `aov` in R uses the name of the factor to identify it.



Relative to the
within group variation …

$$MSW = \frac{\sum_i^K (n_i - 1) S_i^2}{N - K}$$

NOTE: MSW: mean squares within, since it is the sum of squares within groups divided by its degrees of freedom. Note that `aov` in R uses “Residuals” to identify it.

One-Way ANOVA

- Compare the **between** group variation with the **within** group variation using a ratio:

$$F = \frac{MSB}{MSW}; \quad F \sim \mathcal{F}(K-1, N-K)$$

- If the null hypothesis is true (i.e., under H_0):
  - the (population) expected tax value of all properties are equal, regardless of age
  - based on *the sample* observations (random sample), the average tax values (*sample means*) are **not** exactly the same

```
> dat.small %>% group_by(age_factor) %>% summarize(avg_assessment=mean(assessment_k))
# A tibble: 3 x 2
  age_factor avg_assessment
      <fctr>          <dbl>
1          C         521.00
2          M         428.80
3          O         435.14
```

ANOVA

- Study the effect of one or more ***qualitative variables (factors)*** on a ***quantitative response variable***:
  - **Quantitative response**: property tax assessment
  - **Factor**: Age (3 levels: contemporary, modern, old)
- **One-way ANOVA**: 1 factor
- **Two-way ANOVA**: 2 factors
  - **2 factor**: Age and Garage
- t-test is a special case of ANOVA
  - **1 factor with 2 levels**: 2 age periods (e.g., M vs C)

ANOVA table

```
#ANOVA table
summary(aov(assessment_k~age_factor,data=dat.small))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_factor    2  35867   17934   1.226    0.3
## Residuals    67 980328   14632
```

ratio of means of square errors

Between groups variation

Within group variation

Is there enough evidence to claim that the assessment value does not depend on the age of the property?

The F-statistics

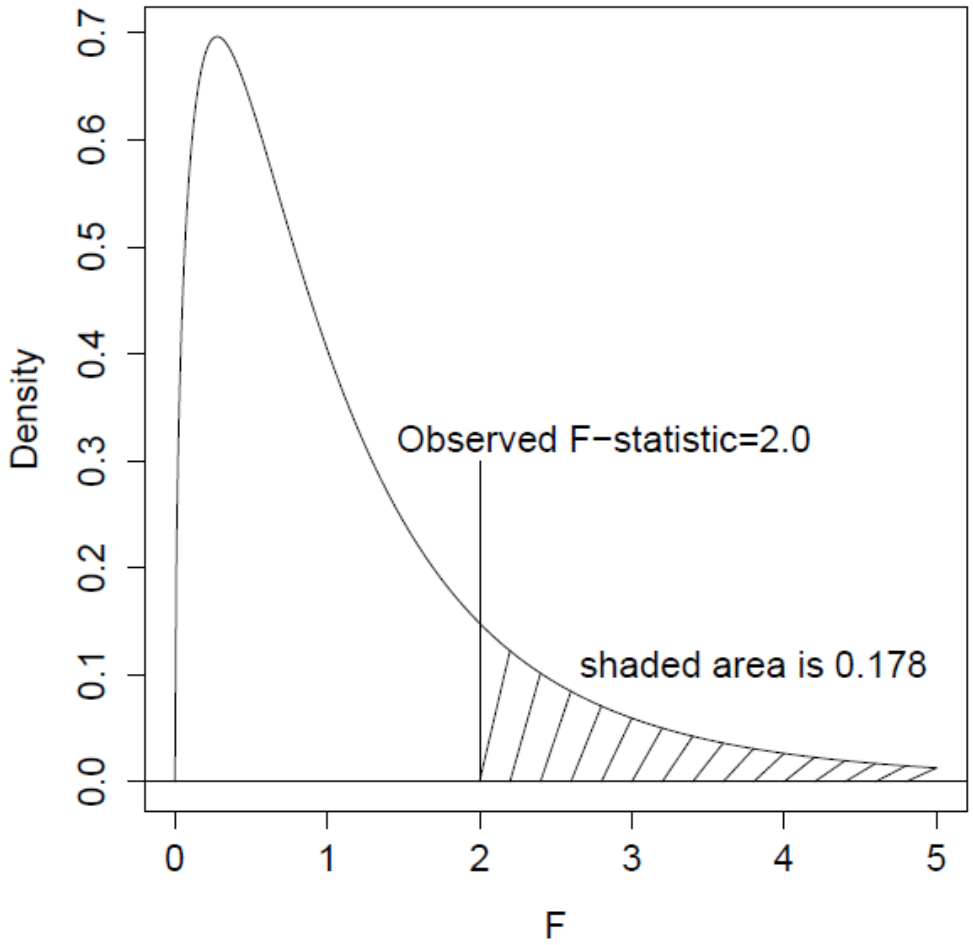


Figure from *www.stat.cmu.edu/~hseltman/309/Book/**chapter7**.pdf*

Right or wrong? Work in groups (5 min)

- ANOVA tests the null hypothesis that the sample means are all equal.
- We use ANOVA to compare the variances of the population.
- A one-way ANOVA is equivalent to a t-test when there are 2 groups to be compared.
- In rejecting the null hypothesis, one can conclude that all the means are different from one another.
- In ANOVA, we assume that all the group sample variances are equal.

- ANOVA tests the null hypothesis that the sample means are all equal.

No, it test the equality of the population means

- We use ANOVA to compare the variances of the population.

No, we use ANOVA to compare the population means

- A one-way ANOVA is equivalent to a t-test when there are 2 groups to be compared.

Correct if we assume that the population variances are equal, 2 groups can be represented as a factor with 2 levels

- In rejecting the null hypothesis, one can conclude that all the means are different from one another.

No, there are at least 2 different population means, we conclude that they are not all equal

- In ANOVA, we assume that all the group sample variances are equal.

No, we assume that the population variance of all groups are equal

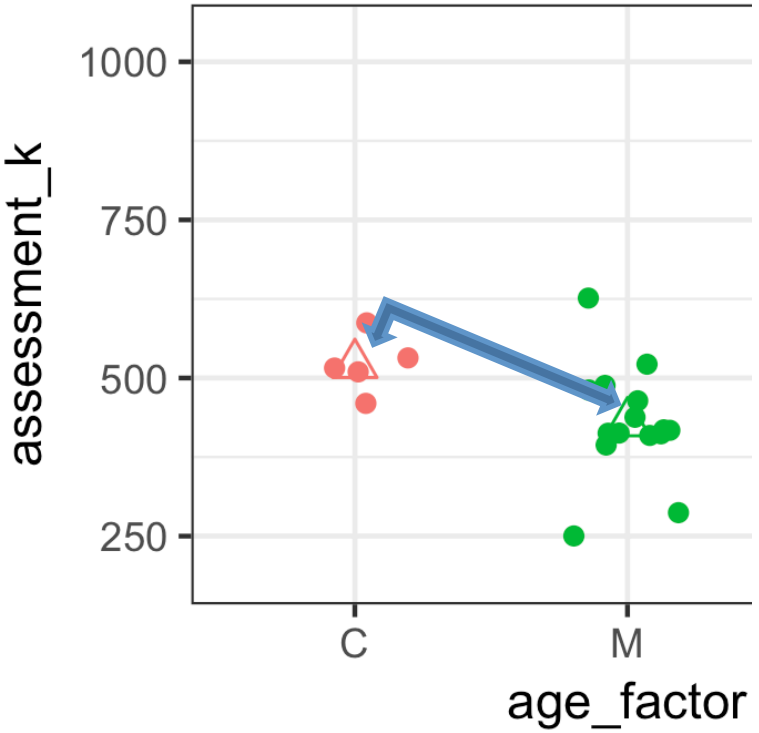
t-test as a special case of ANOVA

Y_i : assessment value for properties built in period i

$$H_0 : \mu_C = \mu_M$$

Two-sample t-test

$$t = \frac{\bar{Y}_M - \bar{Y}_C}{S_p\sqrt{\frac{1}{n_M} + \frac{1}{n_C}}};\ S_p^2 = \frac{\sum_i^2 (n_i - 1)S_i^2}{n_M + n_C - 2}$$



ANOVA: F-test

$$F = \frac{(N-K)\sum_i^K n_i(\bar{Y}_i - \bar{Y})^2}{(K-1)\sum_i^K (n_i - 1)S_i^2}$$

$$\text{for } i = 2,\ t^2 = F$$

```r
#t-test vs ANOVA
#responses within each group
tax.M <-dat.small %>% subset(age_factor =="M", select=assessment_k)
tax.C <-dat.small %>% subset(age_factor =="C", select=assessment_k)

t.test(tax.M,tax.C,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  tax.M and tax.C
## t = -2.2034, df = 18, p-value = 0.04083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -180.111457   -4.288543
## sample estimates:
## mean of x mean of y
##     428.8     521.0
```

$(-2.2034)^2 = 4.855$

```r
#subset of 2 age periods
summary(aov(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_factor    1  31878   31878   4.855 0.0408 *
## Residuals    18 118188    6566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

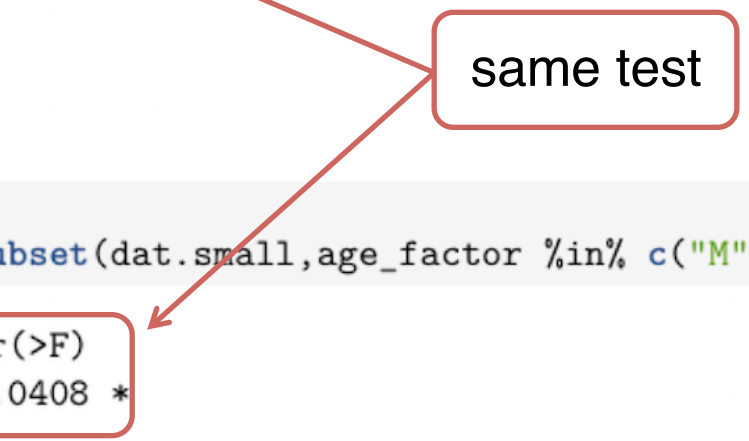
```r
#t-test vs ANOVA
#responses within each group
tax.M <-dat.small %>% subset(age_factor =="M", select=assessment_k)
tax.C <-dat.small %>% subset(age_factor =="C", select=assessment_k)

t.test(tax.M,tax.C,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  tax.M and tax.C
## t = -2.2034, df = 18, p-value = 0.04083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -180.111457   -4.288543
## sample estimates:
## mean of x mean of y
##     428.8     521.0
```

```r
#subset of 2 age periods
summary(aov(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_factor    1  31878   31878   4.855 0.0408 *
## Residuals    18 118188    6566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```r
#t-test vs ANOVA
#responses within each group
tax.M <-dat.small %>% subset(age_factor =="M", select=assessment_k)
tax.C <-dat.small %>% subset(age_factor =="C", select=assessment_k)

t.test(tax.M,tax.C,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  tax.M and tax.C
## t = -2.2034, df = 18, p-value = 0.04083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -180.111457   -4.288543
## sample estimates:
## mean of x mean of y
##     428.8     521.0
```

```r
#subset of 2 age periods
summary(aov(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_factor    1  31878   31878   4.855 0.0408 *
## Residuals    18 118188    6566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Why ANOVA in a Regression course?

Two groups

```
#Linear regression

#LM with 2 age periods
summary(lm(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = subset(dat.small,
##     age_factor %in% c("M", "C")))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.80  -17.55  -10.90   39.45  197.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      36.24  14.377 2.62e-11 ***
## age_factorM   -92.20      41.84  -2.203   0.0408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.03 on 18 degrees of freedom
## Multiple R-squared:  0.2124, Adjusted R-squared:  0.1687
## F-statistic: 4.855 on 1 and 18 DF,  p-value: 0.04083
```

same as t-test

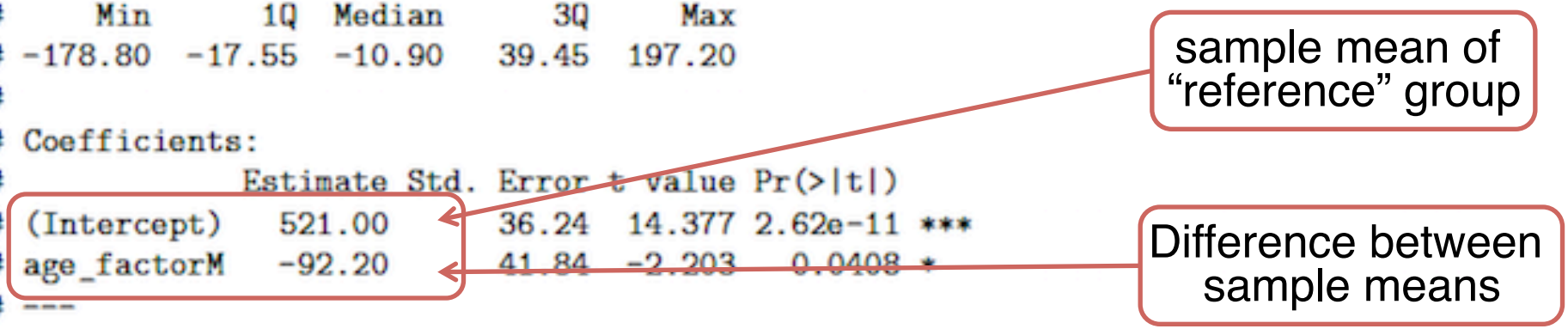
same as ANOVA

Two groups

```
#Linear regression

#LM with 2 age periods
summary(lm(assessment_k~age_factor,data=subset(dat.small,age_factor %in% c("M","C"))))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = subset(dat.small,
##     age_factor %in% c("M", "C")))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.80  -17.55  -10.90   39.45  197.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      36.24  14.377 2.62e-11 ***
## age_factorM   -92.20      41.84  -2.203   0.0408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.03 on 18 degrees of freedom
## Multiple R-squared:  0.2124, Adjusted R-squared:  0.1687
## F-statistic: 4.855 on 1 and 18 DF,  p-value: 0.04083
```



```r
#More than 2 groups

#LM with 3 age periods
summary(lm(assessment_k~age_factor,data=dat.small))
```

```
##
## Call:
## lm(formula = assessment_k ~ age_factor, data = dat.small)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -250.14  -74.89  -16.97   51.36  612.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   521.00      54.10   9.631 2.87e-14 ***
## age_factorM   -92.20      62.46  -1.476    0.145
## age_factor0   -85.86      56.74  -1.513    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121 on 67 degrees of freedom
## Multiple R-squared:  0.0353, Adjusted R-squared:  0.006498
## F-statistic: 1.226 on 2 and 67 DF,  p-value: 0.3001
```

different from t-test!!

same test

```r
#ANOVA with 3 age periods
summary(aov(assessment_k~age_factor,data=dat.small))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_factor    2  35867   17934   1.226    0.3
## Residuals    67 980328   14632
```