431 Class 18

Thomas E. Love

2017-10-31

Gelman on Statistical Significance

"... we use the term statistically significant in the conventional way, to mean that an estimate is **at least two standard errors away** from some"null hypothesis" or prespecified value that would indicate no effect present. An estimate is statistically insignificant if the observed value could reasonably be explained by simple chance variation, much in the way that a sequence of 20 coin tosses might happen to come up 8 heads and 12 tails; we would say that this result is not statistically significantly different from chance. More precisely, the observed proportion of heads is 40 percent but with a standard error of 11 percent - thus, the data are less than two standard errors away from the null hypothesis of 50 percent, and the outcome could clearly have occurred by chance. Standard error is a measure of the variation in an estimate and gets smaller as a sample size gets larger, converging on zero as the sample increases in size."

Gelman's blog (2017-10-28)

Today's Agenda

- Discussion of the Class 17 In-Class Survey
- Comparing More than Two Populations: The Analysis of Variance
- Pairwise Comparisons of Means after a Significant ANOVA
  - Multiple Comparisons
  - Bonferroni and Tukey HSD approaches
- More on p Values and Statistical Significance

Today's R Setup

```
library(forcats); library(tidyverse)

source("Love-boost.R")
dm192 <- read.csv("data/dm192.csv") %>% tbl_df
class17a <- read.csv("data/class17a.csv") %>% tbl_df
class17b <- read.csv("data/class17b.csv") %>% tbl_df
```

In-Class Survey from Class 17

In-Class Survey (`class17a` data)

We chose (using a computer) a random number between 0 and 100.

Your number is X = 10 (or 65).

1. Do you think the percentage of countries which are in Africa, among all those in the United Nations, is higher or lower than X?
2. Give your best estimate of the percentage of countries which are in Africa, among all those in the United Nations.

The facts

- There are 193 sovereign states that are members of the UN.
- The African regional group has 54 member states, so that's 28%.
- UN regions for countries are this Wikipedia link
- The `class17a` data set contains the answers to these questions from 185 students asked the same questions in the same way over the past four years (since 2014).

A troubling situation

We chose (using a computer) a random number between 0 and 100. Your number is X = 65.

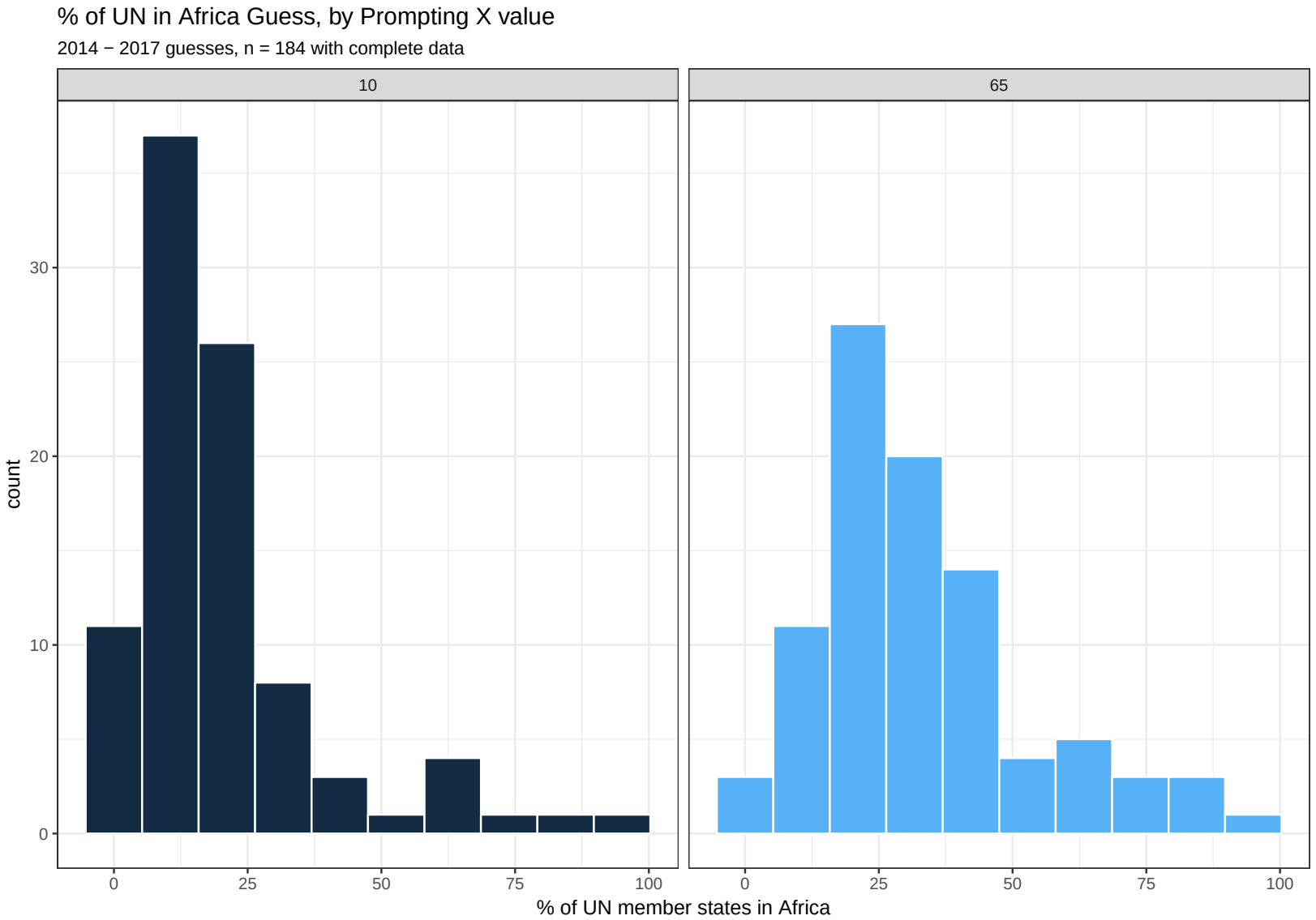
1. Do you think the *percentage* of countries which are in Africa, among all those in the United Nations, is **higher** or **lower** than X?

   Circle your answer: HIGHER than X LOWER than X

2. Give your best estimate of the *percentage* of countries which are in Africa, among all those in the United Nations.

   My Answer: 20 percent.

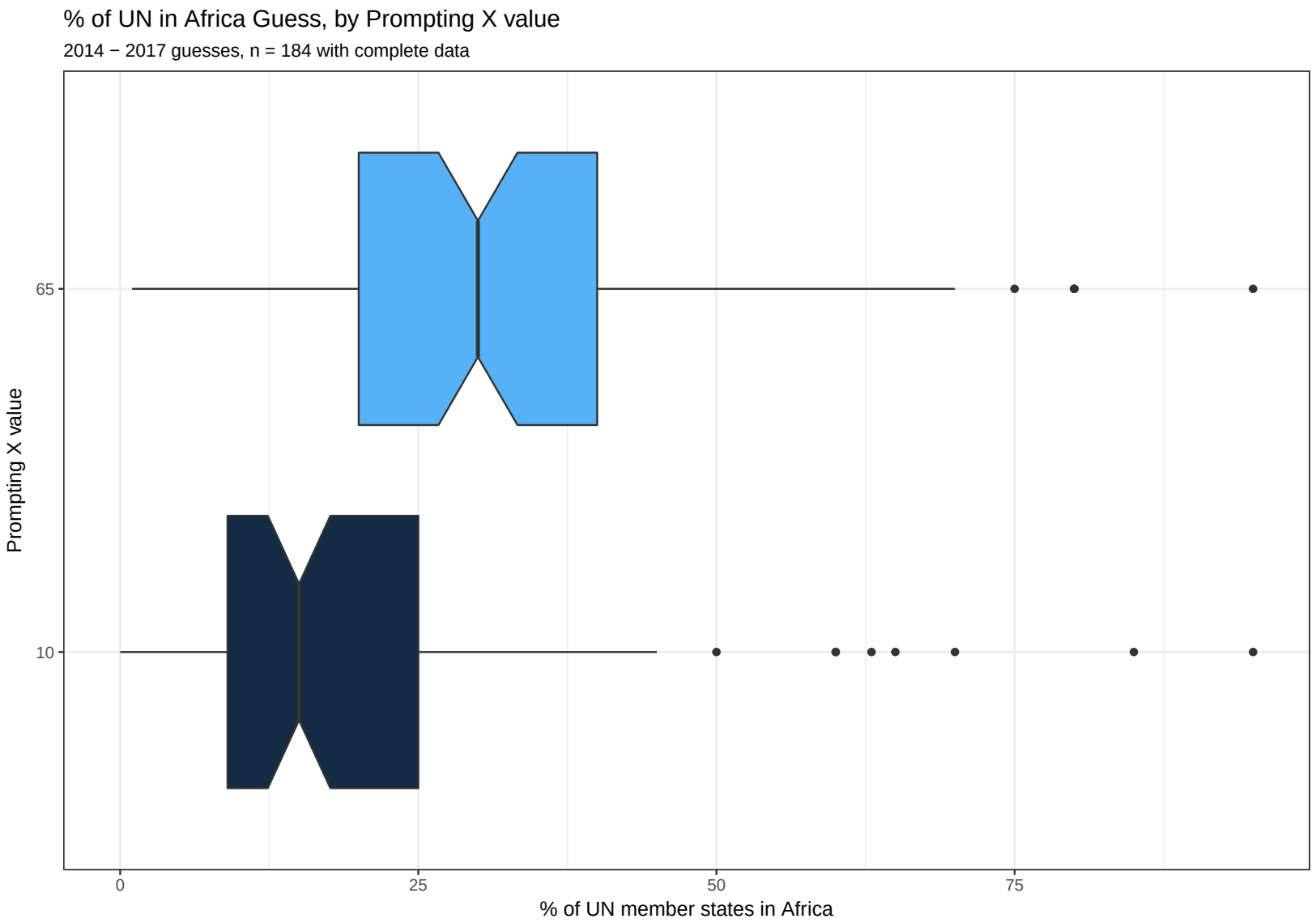
class17a Africa percentage guess by X = 10 or 65



`class17a` Analysis, Step-by-Step

1. What is the outcome under study?
2. What are the (in this case, two) treatment/exposure groups?
3. Were the data collected using matched / paired samples or independent samples?
4. Are the data a random sample from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the sample to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing/confidence interval generation?
7. If we have paired samples, did pairing help reduce nuisance variation?
8. If we have paired samples, what does the distribution of sample paired differences tell us about which inferential procedure to use?
9. If we have independent samples, what does the distribution of each individual sample tell us about which inferential procedure to use?

`class17a` Africa percentage guess by X = 10 or 65



class17a Descriptive Statistics

```
class17a %>%
  filter(!is.na(africa.pct)) %>%
  group_by(x.value) %>%
  summarise(n(), mean(africa.pct),
            sd = round(sd(africa.pct),2),
            median = median(africa.pct))
```

```
# A tibble: 2 x 5
  x.value `n()` `mean(africa.pct)`    sd median
    <int> <int>              <dbl> <dbl>  <int>
1      10    93           20.96774 17.63     15
2      65    91           33.00000 19.29     30
```

class17a comparisons (results: next slide)

```
t.test(africa.pct ~ x.value,
       data = class17a) # Welch
t.test(africa.pct ~ x.value, data = class17a,
       var.equal = TRUE) # Pooled t
wilcox.test(africa.pct ~ x.value, conf.int = TRUE,
            data = class17a)
set.seed(43123)
bootdif(class17a$africa.pct, class17a$x.value)
```

class17a Comparing Two Populations

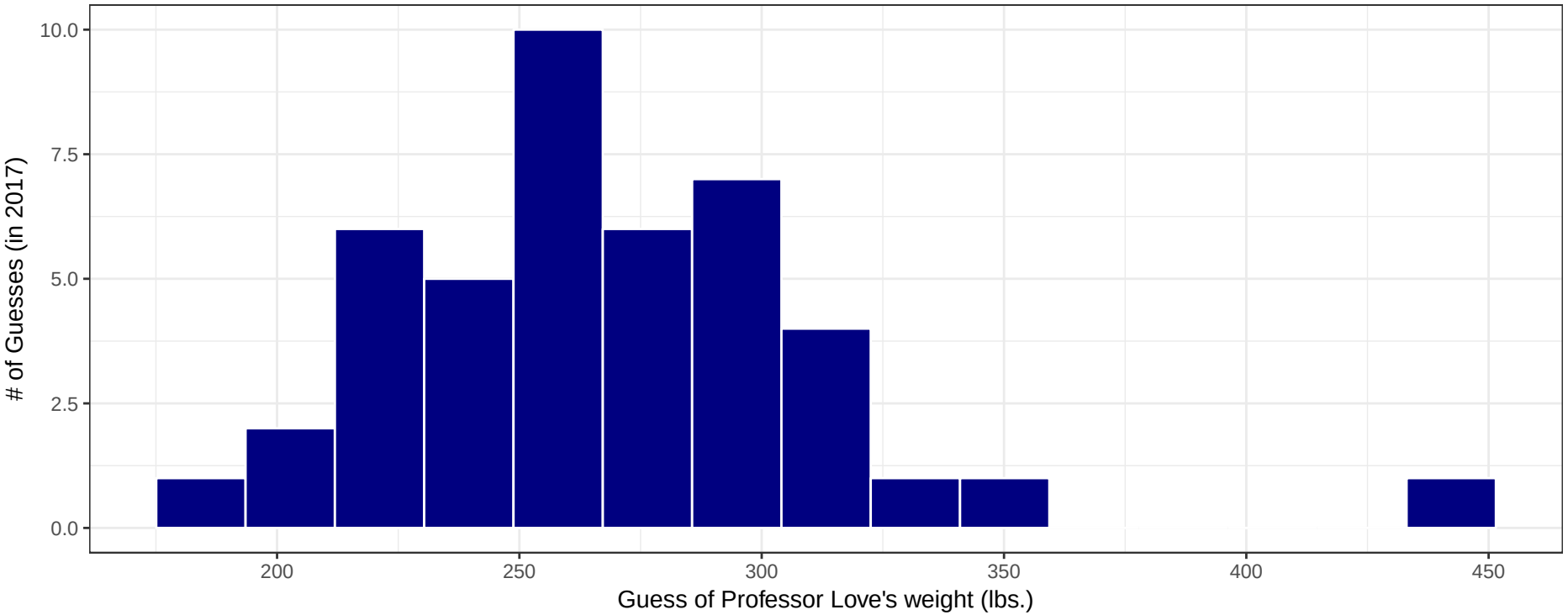
$$\Delta = \mu_{65} - \mu_{10}$$

Procedure	Est. Δ	95% CI for Δ	p
Welch t	12.0	(6.7, 17.4)	1.75e-05
Pooled t	12.0	(6.7, 17.4)	1.71e-05
Rank Sum	12.0	(8.0, 15.0)	6.06e-08
Bootstrap	12.0	(6.6, 17.5)	$< .05$

Conclusions?

In-Class Survey (`class17b` data)

3. Provide a point estimate for Dr. Love's current weight (in pounds.)



- 2017 Weight Guesses: $n = 44$, $\bar{x} = 267.7$ lbs., $s = 46.5$ lbs.
- Five Number Summary: 182 240 260 293 440

50% and 90% "Intervals" from Group Estimates

4. Now estimate one interval, which you believe has a 50% chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a 90% interval.

We have $n = 44$ independent guesses, with $\bar{x} = 267.7$ lbs., $s = 46.5$ lbs. Let's first obtain quantiles, and use the crowd's wisdom.

```
quantile(class17b$love.lbs,
  probs = c(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95))
```

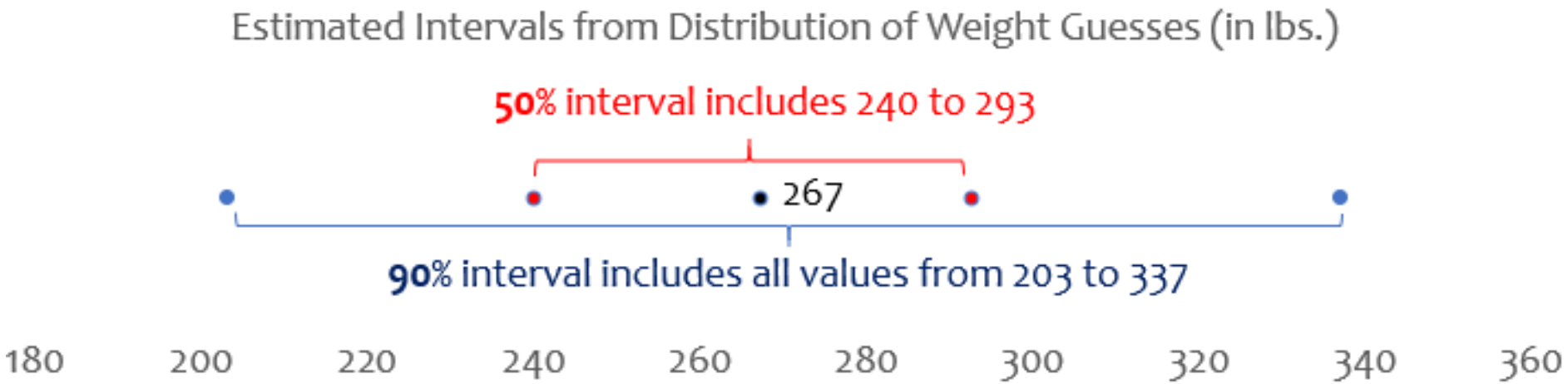
```
   5%    10%    25%    50%    75%    90%    95%
203.0 220.0 240.0 260.0 292.5 320.0 337.0
```

- What's a rational 50% interval for estimating my weight?
- How about a 90% interval?

One Possible, Rational, Set of Intervals

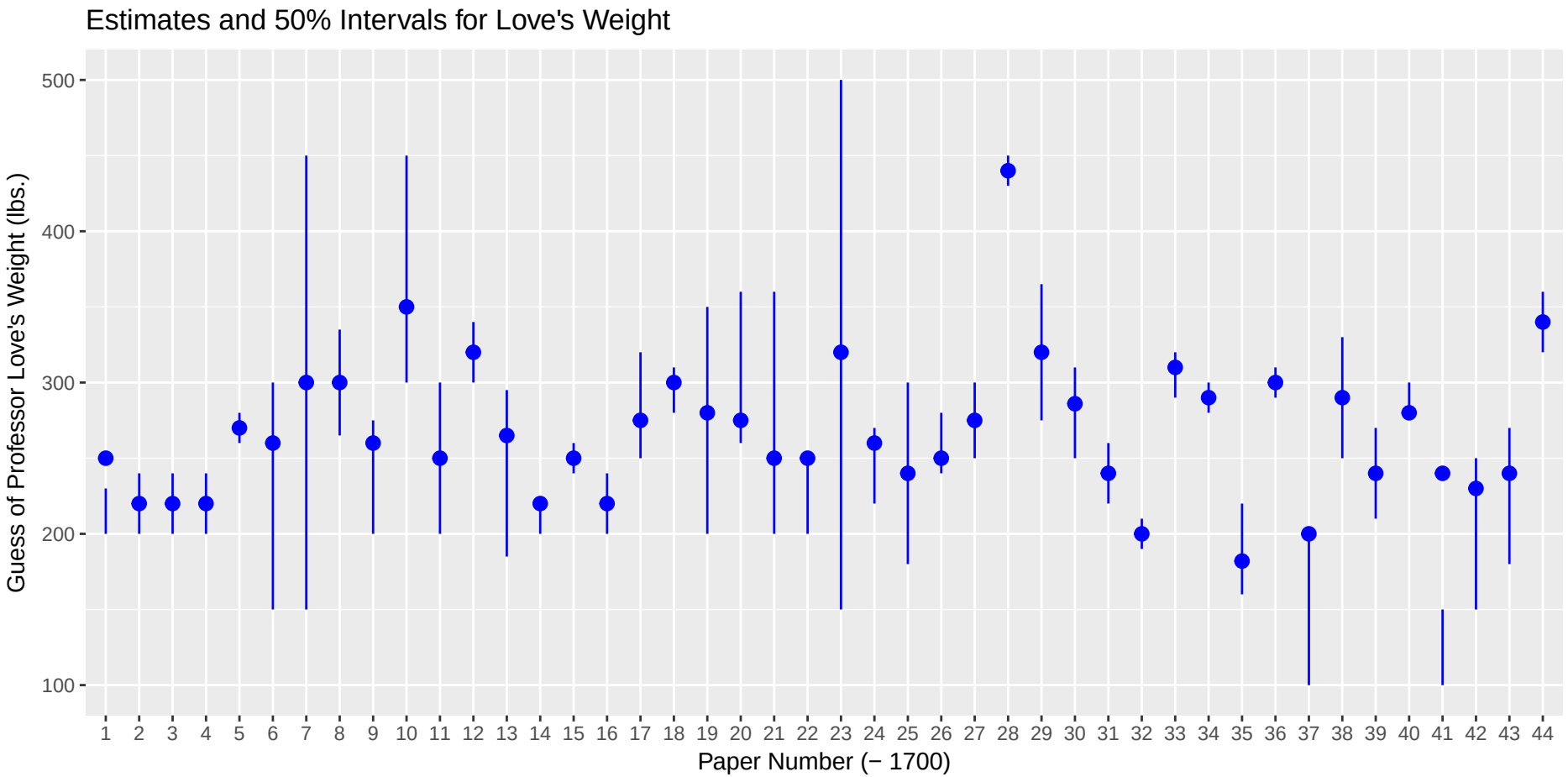
Suppose my estimate is 267 pounds.

- Then suppose I assign probability 0.50 to the interval (240, 293)
- And suppose I assign probability 0.90 to the interval (203, 337)



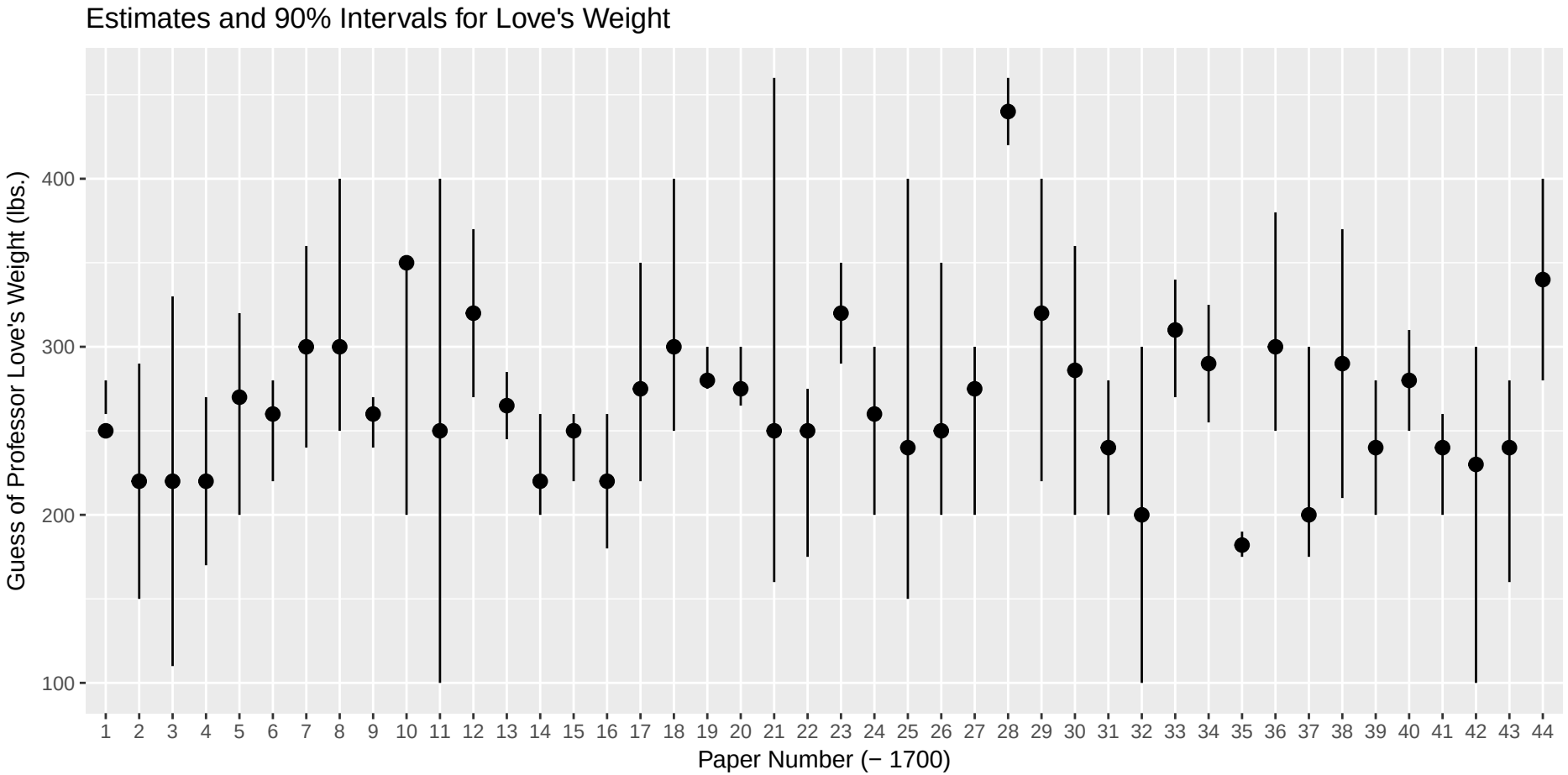
In-Class Survey (`class17b` data)

4a. Now estimate one interval, which you believe has a 50% chance of including Dr. Love's current weight (again, in pounds.)

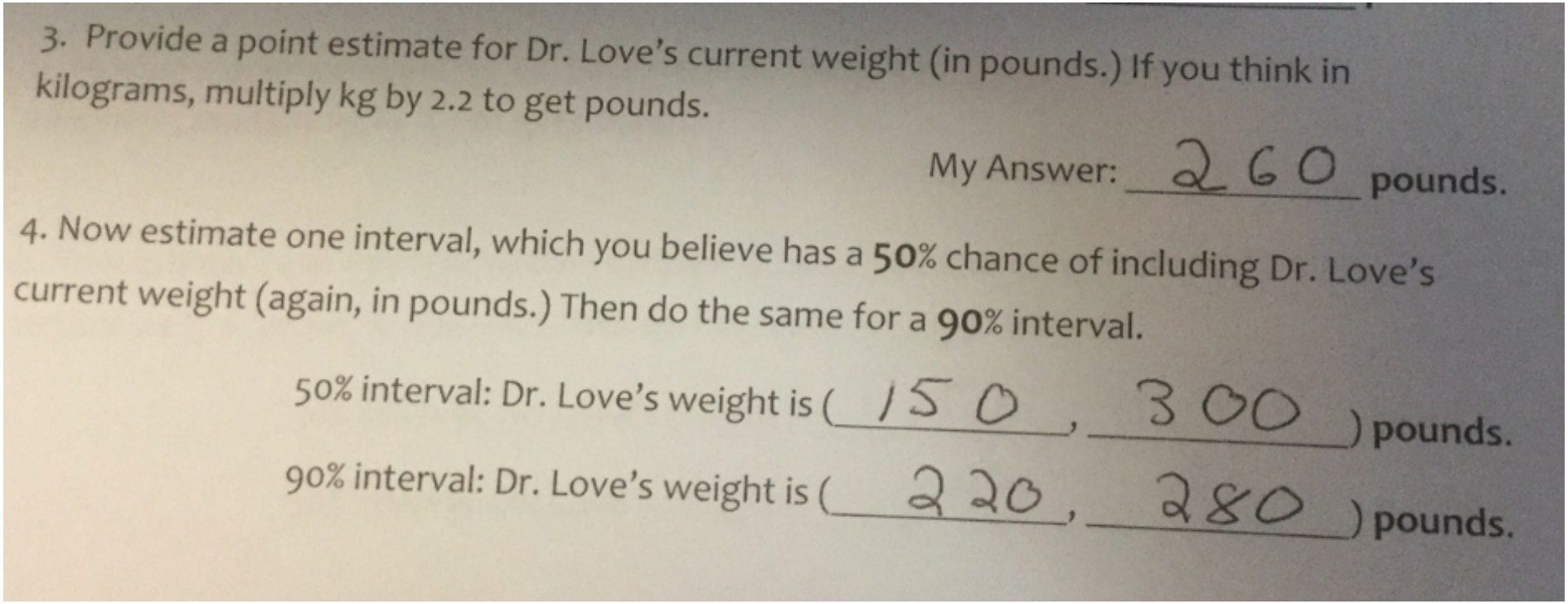


In-Class Survey (`class17b` data)

4b. Now do the same, but for a 90% interval. . .



Some Troubling Selections, 1



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 260 pounds.

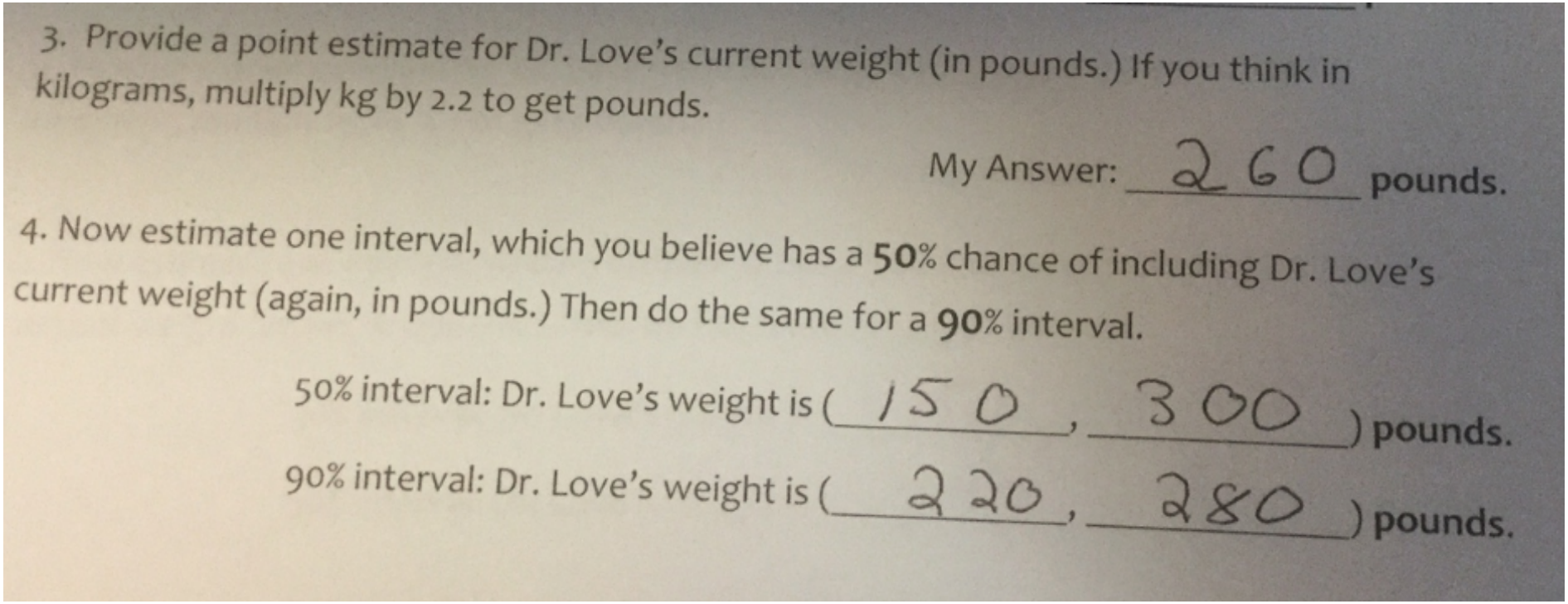
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (150 , 300) pounds.

90% interval: Dr. Love's weight is (220 , 280) pounds.

- Why does this set of intervals not make sense?

Some Troubling Selections, 1



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 260 pounds.

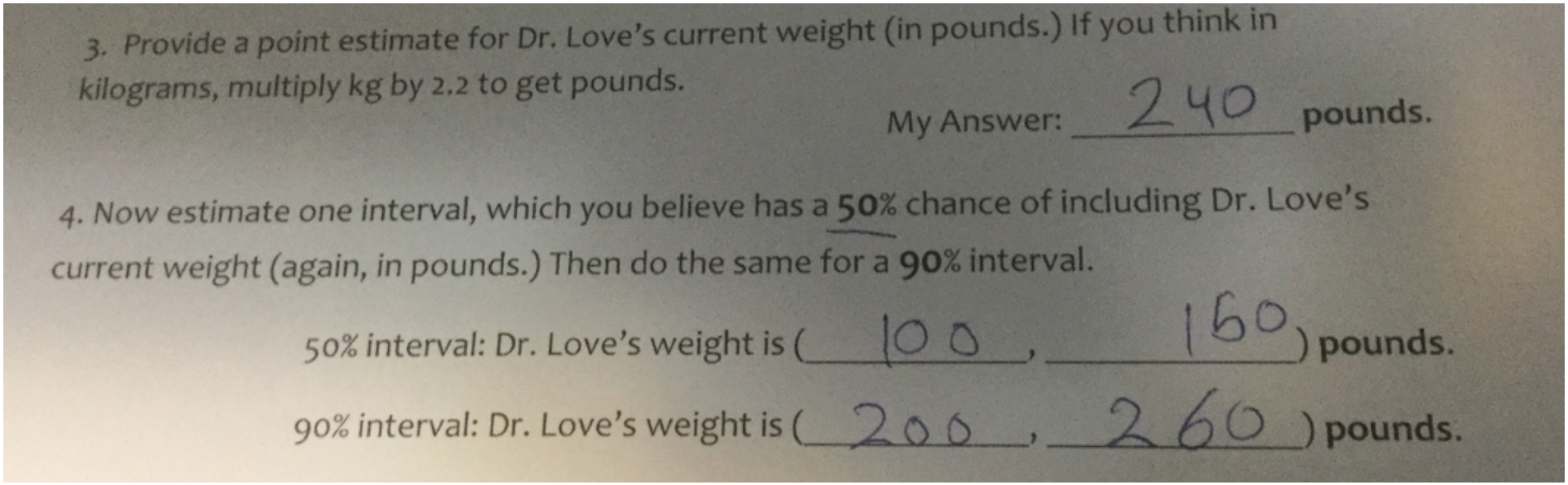
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (150 , 300) pounds.

90% interval: Dr. Love's weight is (220 , 280) pounds.

- Why does this set of intervals not make sense?
- There were **10** students (out of 43) who had a wider 50% interval than 90% interval.

Some Troubling Selections, 2



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 240 pounds.

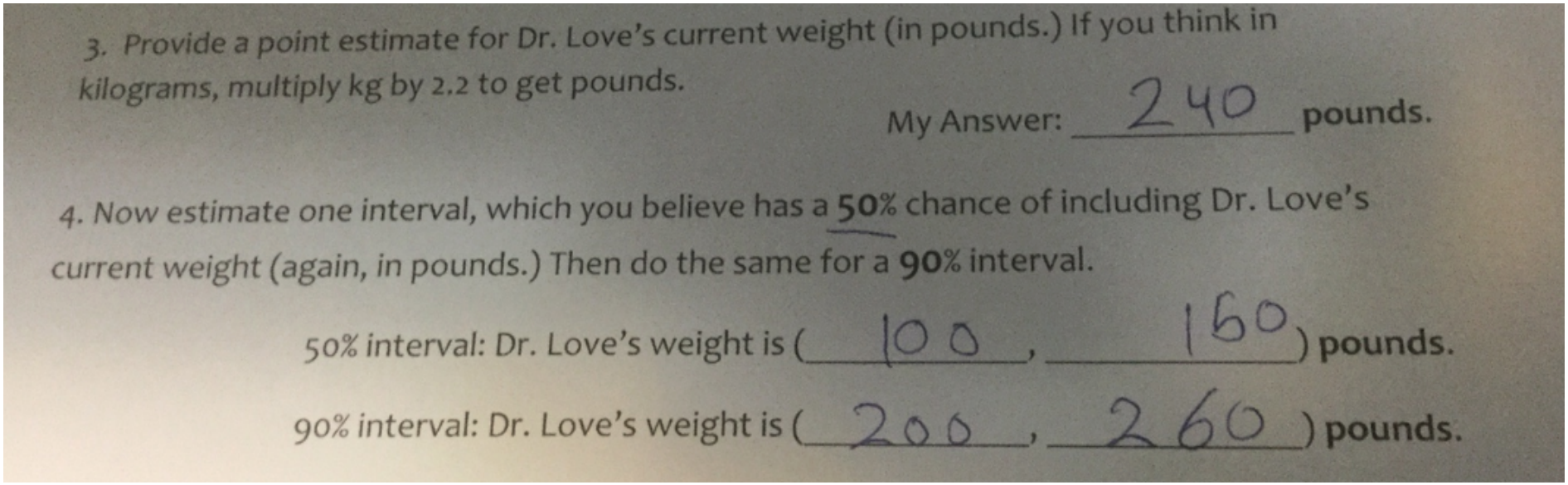
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (100 , 160) pounds.

90% interval: Dr. Love's weight is (200 , 260) pounds.

- It wasn't clear enough that the interval estimate was meant to surround the point estimate.

Some Troubling Selections, 2



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 240 pounds.

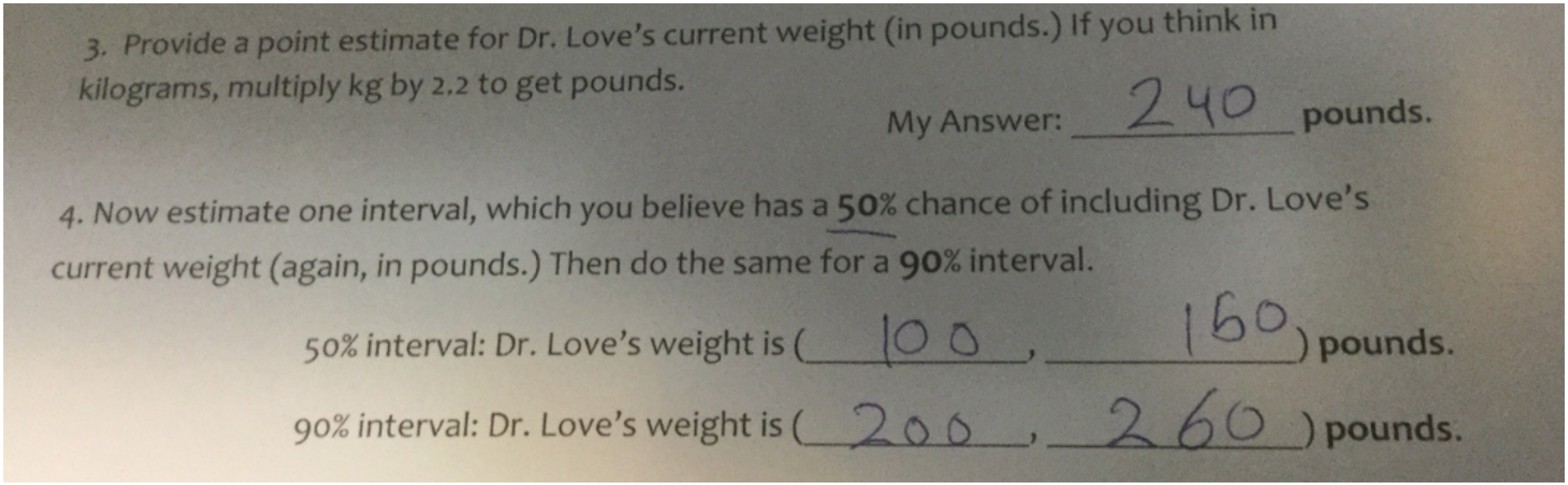
4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

50% interval: Dr. Love's weight is (100 , 160) pounds.

90% interval: Dr. Love's weight is (200 , 260) pounds.

- It wasn't clear enough that the interval estimate was meant to surround the point estimate.
- There were **6** students out of 43 with this problem in their 50% interval, **2** in their 90% interval.

Some Troubling Selections, 2



3. Provide a point estimate for Dr. Love's current weight (in pounds.) If you think in kilograms, multiply kg by 2.2 to get pounds.

My Answer: 240 pounds.

4. Now estimate one interval, which you believe has a **50%** chance of including Dr. Love's current weight (again, in pounds.) Then do the same for a **90%** interval.

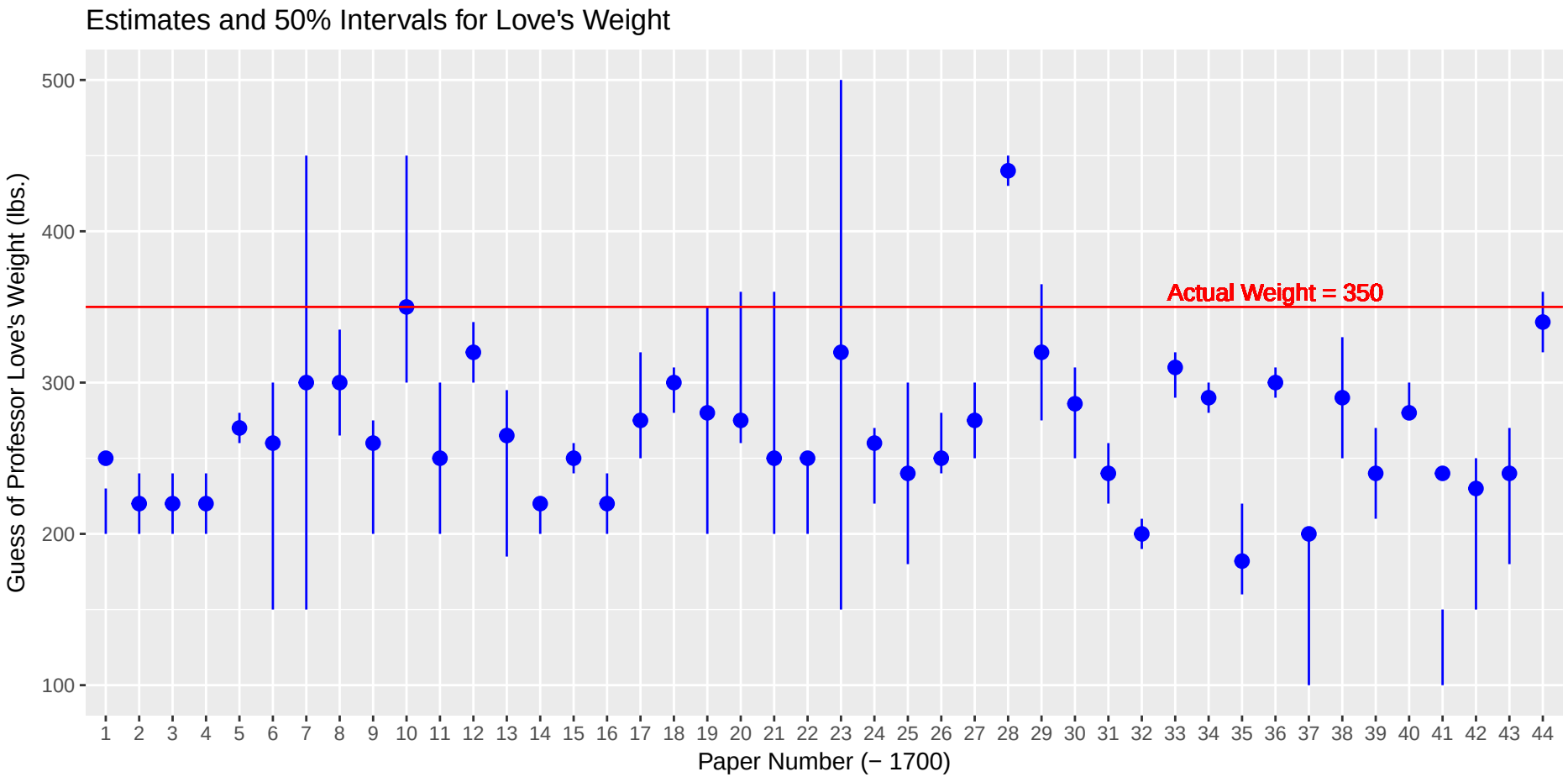
50% interval: Dr. Love's weight is (100 , 160) pounds.

90% interval: Dr. Love's weight is (200 , 260) pounds.

- It wasn't clear enough that the interval estimate was meant to surround the point estimate.
- There were **6** students out of 43 with this problem in their 50% interval, **2** in their 90% interval.
- For **15** students, the 90% interval did not contain the 50% interval.

The facts (with 50% intervals)

On 2017-10-26, Dr. Love actually weighed 350 lbs. or 158.8 kg or 25 stone, dressed but without shoes.



- **8** of the 43 50% intervals estimated by students included 350 lbs.

The facts (with 90% intervals)

On 2017-10-26, Dr. Love actually weighed 350 lbs.

- **16** of the 43 50% intervals estimated by students included 350 lbs.

Analysis of Variance (Section 28, Course Notes)

Analysis of Variance to Compare More Than Two Population Means using Independent Samples

Suppose we want to compare more than two population means, and we have collected three or more independent samples.

This is analysis of a continuous outcome variable on the basis of a single categorical factor — in fact, it's often called **one-factor** ANOVA or **one-way** ANOVA to indicate that the outcome is being split up into the groups defined by a single factor.

- H_0: population means in each group are the same
- H_A: H_0 isn't true; at least one μ differs from the others

When there are just two groups, then this boils down to an F test that is equivalent to the Pooled t test.

One-Way ANOVA

If we have a grouping factor with k levels, then we are testing:

- H_0: $\mu_1 = \mu_2 = ... = \mu_k$ vs.
- H_A: At least one of the population means $\mu_1, \mu_2, ..., \mu_k$ is different from the others.

Today's Example

We'll look at the `dm192` data again.

- Outcome is the `a1c` value (measured as a percentage),
- Factor is the insurance group (we'll compare 3 categories).

The dm192 data: Comparing Insurance Groups on Hemoglobin A1c

```
dm.ins <- select(dm192, pt.id, insurance, a1c)
summary(dm.ins)
```

```
     pt.id              insurance       a1c       
 Min.   :  1.00   commercial:39   Min.   : 5.400  
 1st Qu.: 48.75   medicaid  :67   1st Qu.: 6.300  
 Median : 96.50   medicare  :80   Median : 7.300  
 Mean   : 96.50   uninsured : 6   Mean   : 7.973  
 3rd Qu.:144.25                   3rd Qu.: 9.000  
 Max.   :192.00                   Max.   :16.100  
                                  NA's   :4       
```

- For now, we'll collapse the 6 uninsured in with Medicaid patients, and we'll drop the four cases without an A1c value.

Collapse medicaid and uninsured, drop missing a1c

```
dm.ins <- dm.ins %>%
    mutate(ins.3cat = fct_recode(insurance,
                "Commercial" = "commercial",
                "Medicare" = "medicare",
                "Medicaid/Unins." = "medicaid",
                "Medicaid/Unins." = "uninsured")) %>%
  filter(!is.na(a1c))
```

Summarize A1c by Insurance (3 categories)

```
dm.ins %>%
  group_by(ins.3cat) %>%
  summarise(n = n(), mean = round(mean(a1c),2),
            sd = round(sd(a1c),2), median = median(a1c))
```

```
# A tibble: 3 x 5
         ins.3cat     n  mean    sd median
           <fctr> <int> <dbl> <dbl>  <dbl>
1      Commercial    39  8.10  2.03    7.6
2 Medicaid/Unins.    73  8.12  2.35    7.5
3        Medicare    76  7.76  2.26    7.0
```

One-Way ANOVA for the dm.ins Data

If we have a grouping factor (insurance) with *3* levels, then we are testing:

- H_0: $\mu_{Comm.} = \mu_{Medicare} = \mu_{Medicaid/Unins.}$ vs.
- H_A: At least one of the population means is different from the others.

```
anova(lm(a1c ~ ins.3cat, data = dm.ins))
```

```
Analysis of Variance Table

Response: a1c
           Df Sum Sq Mean Sq F value Pr(>F)
ins.3cat    2   5.55  2.7763  0.5466 0.5798
Residuals 185 939.60  5.0789
```

Elements of the ANOVA Table

The ANOVA table breaks down the variation in the outcome explained by the k levels of the factor of interest, and the variation in the outcome which remains (the Residual, or Error).

```
Analysis of Variance Table

Response: a1c
           Df Sum Sq Mean Sq F value Pr(>F)
ins.3cat    2   5.55  2.7763  0.5466 0.5798
Residuals 185 939.60  5.0789
```

- Df = degrees of freedom, Sum Sq = Sum of Squares,
- Mean Sq = Mean Square (Sum of Squares / df)
- F value = F test statistic, Pr(>F) = p value

The Degrees of Freedom

```
            Df
ins.3cat     2
Residuals  185
```

- The **degrees of freedom** attributable to the factor of interest (here, `ins.3cat`) is the number of levels of the factor minus 1.
  - Here, we have three insurance category levels, so df(`ins.3cat`) = 2.
- The total degrees of freedom are the number of observations (across all levels of the factor) minus 1.
  - We have 188 patients left in our `dm.ins` study after removing the four with missing A1c, so df(Total) = 187, although the Total row isn't shown here.
- Residual df = Total df - Factor df = 187 - 2 = 185.

The Sums of Squares

```
            Df Sum Sq
ins.3cat     2   5.55
Residuals  185 939.60
```

- The **sum of squares** (SS) represents variation explained.
- SS(Factor) is the sum across all levels of the factor of the sample size for the level multiplied by the squared difference between the level mean and the overall mean across all levels. SS(`ins.3cat`) = 5.55
- SS(Total) = sum across all observations of the square of the difference between the individual values and the overall mean.
  - Here SS(Total) = 5.55 + 939.60 = 945.15
- Residual SS = Total SS – Factor SS.

η^2, the Proportion of Variation Explained by ANOVA

```
            Df Sum Sq
ins.3cat     2   5.55
Residuals  185 939.60
```

- η^2 ("eta-squared") is equivalent to R^2 in a linear model.
  - η^2 = SS(Factor) / SS(Total) = the proportion of variation in our outcome (here, hemoglobin A1c) explained by the variation between levels of our factor (here, our three insurance groups)
  - In our case, $\eta^2 = 5.55 / (5.55 + 939.60) = 5.55 / 945.15 = 0.0059$
- So, insurance group accounts for about 0.59% of the variation in hemoglobin A1c observed in these data.

The Mean Square

```
            Df Sum Sq Mean Sq
ins.3cat     2   5.55  2.7763
Residuals  185 939.60  5.0789
```

- The Mean Square is the Sum of Squares divided by the degrees of freedom, so MS(Factor) = SS(Factor)/df(Factor).
- MS(ins.3cat) = SS(ins.3cat)/df(ins.3cat) = 5.55 / 2 = 2.78.
- MS(Residuals) = SS(Residuals) / df(Residuals) = 939.60 / 185 = 5.08.
  - MS(Residuals) estimates the residual variance, corresponds to σ^2 in the underlying linear model
  - MS(Residuals) = 5.0789, so Residual standard error = $\sqrt{5.0789}$ = 2.25 percentage points.

The F Test Statistic and p Value

```
Analysis of Variance Table

Response: a1c
           Df Sum Sq Mean Sq F value Pr(>F)
ins.3cat    2   5.55  2.7763  0.5466 0.5798
Residuals 185 939.60  5.0789
```

- F value = MS(ins.3cat) / MS(Residuals) = 2.78 / 5.08 = 0.55
- For an F distribution with 2 and 185 degrees of freedom, this F value yields $p = 0.58$

What is our conclusion regarding our test of our ANOVA hypotheses?

- H_0: $\mu_{Commercial} = \mu_{MedicaidorUninsured} = \mu_{Medicare}$ vs.
- H_A: H_0 is not true

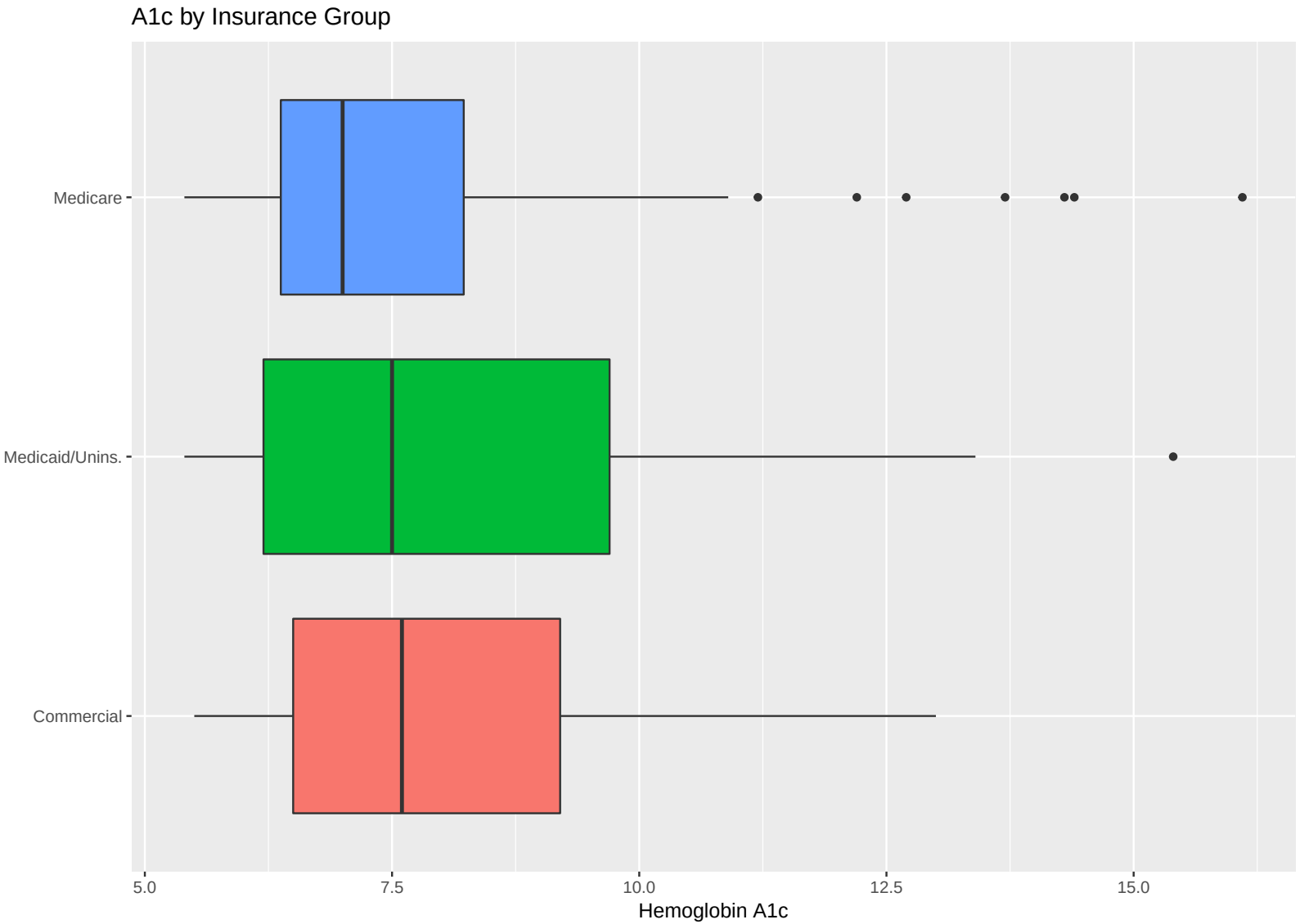
ANOVA Assumptions

The assumptions behind analysis of variance are the same as those behind a linear model. Of specific interest are:

- The samples obtained from each group are independent.
- Ideally, the samples from each group are a random sample from the population described by that group.
- In the population, the variance of the outcome in each group is equal. (This is less of an issue if our study involves a balanced design.)
- In the population, we have Normal distributions of the outcome in each group.

Happily, the F test is fairly robust to violations of the Normality assumption.

Can we assume population A1c levels are Normal?



Non-Parametric Alternative: Kruskal-Wallis Test

```
kruskal.test(a1c ~ ins.3cat, data = dm.ins)
```

```
    Kruskal-Wallis rank sum test

data:  a1c by ins.3cat
Kruskal-Wallis chi-squared = 1.7809, df = 2,
p-value = 0.4105
```

Rank Sum test for

- H_0: Center of Commercial distribution = Center of Medicaid or Uninsured distribution = Center of Medicare distribution vs.
- H_A: H_0 not true.

Another Way to get our ANOVA Results

H_0: H_0: $\mu_{Commercial} = \mu_{MedicaidorUninsured} = \mu_{Medicare}$ vs. H_A: H_0 not true.

```
summary(aov(a1c ~ ins.3cat, data = dm.ins))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
ins.3cat      2    5.6   2.776   0.547   0.58
Residuals   185  939.6   5.079
```

Regression on Indicator Variables = Analysis of Variance

Yet another way to obtain an even more complete analog to the pooled t test is to run a linear regression model to predict the outcome (here, a1c) on the basis of the categorical factor, insurance group. We run the following ...

```
summary(lm(a1c ~ ins.3cat, data = dm.ins))
```

Linear Model Summary Output

```
Call:
lm(formula = a1c ~ ins.3cat, data = dm.ins)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7219 -1.6000 -0.6432  1.0855  8.3355

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.10000    0.36087  22.446   <2e-16 ***
ins.3catMedicaid/Unins.  0.02192    0.44699   0.049    0.961
ins.3catMedicare        -0.33553    0.44391  -0.756    0.451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.254 on 185 degrees of freedom
Multiple R-squared:  0.005875,  Adjusted R-squared:  -0.004872
F-statistic: 0.5466 on 2 and 185 DF,  p-value: 0.5798
```

Linear Model Results

- **Residual standard error: 2.254** on 185 degrees of freedom
- **Multiple R-squared: 0.005875**, Adjusted R-squared: -0.004872
- F-statistic: 0.5466 on 2 and 185 DF, p-value: 0.5798

Indicator Variable Regression

The linear model uses two **indicator variables**, sometimes called **dummy variables**.

- Each takes on the value 1 when its condition is met, and 0 otherwise.
- With three race categories, we need two indicator variables (we always need one fewer indicator than we have levels of the factor).
- Here, we have a baseline category (which is taken to be `Commercial` in this case) and then indicators for `Medicaid or Uninsured` and for `Medicare`.

K-1 indicators specify K categories

These two indicator variables completely specify the insurance category for any subject, as follows:

Insurance Category	`var1`	`var2`
Commercial	0	0
Medicaid/Unins.	1	0
Medicare	0	1

- `var1` is `ins.3catMedicaid/Unins.`
- `var2` is `ins.3catMedicare`

The Regression Equation

What is the regression equation here?

```
Call: lm(formula = a1c ~ ins.3cat, data = dm.ins)

Coefficients              Estimate Std. Err.      t  Pr(>|t|)
(Intercept)                8.10000  0.36087  22.446   <2e-16 ***
ins.3catMedicaid/Unins.    0.02192  0.44699   0.049    0.961
ins.3catMedicare          -0.33553  0.44391  -0.756    0.451
```

Equation specifies the three sample means

- A1c = 8.1 + 0.02 [Medicaid or Uninsured] - 0.34 [Medicare]
- [group] is 1 if the patient is in that group, and 0 otherwise

The Model predictions are Sample Means

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.10000	0.36087	22.446	<2e-16	***
Medicaid/Uninsured	0.02192	0.44699	0.049	0.961	
Medicare	-0.33553	0.44391	-0.756	0.451	

Model Predictions:

- A1c = 8.1 if in the Commercial group
- A1c = 8.1 + 0.02192 = 8.12 if in the Medicaid or Uninsured group
- A1c = 8.1 - 0.33553 = 7.76 if in the Medicare group

K-Sample Study Design, Comparing Means

1. What is the outcome under study?
2. What are the (in this case, $K > 2$) treatment/exposure groups?
3. Were the data in fact collected using independent samples?
4. Are the data random samples from the population(s) of interest? Or is there at least a reasonable argument for generalizing from the samples to the population(s)?
5. What is the significance level (or, the confidence level) we require here?
6. Are we doing one-sided or two-sided testing?
7. What does the distribution of each individual sample tell us about which inferential procedure to use?
8. Are there statistically meaningful differences between population means?
9. If an overall test is significant, can we identify pairwise comparisons of means that show significant differences using an appropriate procedure that protects against Type I error expansion due to multiple comparisons?

A New Comparison using dm192

Let's look at the dm192 data again, but now we'll study dbp (diastolic blood pressure) as our outcome of interest.

- We'll first use ANOVA make a comparison between the four levels of insurance (Medicare, Commercial, Medicaid, Uninsured).
- Later, we'll compare the average dbp across the four practices (A, B, C and D) included in the dm192 sample.

Analysis of Variance for dbp by insurance

H_0: $\mu_{Medicare} = \mu_{Commercial} = \mu_{Medicaid} = \mu_{Uninsured}$ vs. H_A: H_0 not true.

```
summary(aov(dbp ~ insurance, data = dm192))
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
insurance     3   1909   636.2   5.275 0.00163 **
Residuals   188  22672   120.6
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So which of the pairs of means are significantly different?

The Problem of Multiple Comparisons

1. Suppose we compare Medicare to Commercial, using a test with $\alpha = 0.05$
2. Then we compare Medicare to Medicaid on the same outcome, also using $\alpha = 0.05$
3. Then we compare Medicare to Uninsured, also with $\alpha = 0.05$
4. Suppose we compare Commercial to Medicaid with $\alpha = 0.05$
5. Then we compare Commercial to Uninsured with $\alpha = 0.05$
6. Then we compare Medicaid to Uninsured with $\alpha = 0.05$

What is our overall α level across these six comparisons?

The Problem of Multiple Comparisons

What is our overall α level across these six comparisons?

- It could be as bad as 0.05 + 0.05 + 0.05 + 0.05 + 0.05 + 0.05, or 0.30.
- Rather than our nominal 95% confidence, we have something as low as 70% confidence across this set of simultaneous comparisons.
- Does it matter if we *pre-plan* the comparisons or not?

The Bonferroni solution

1. Suppose we compare Medicare to Commercial, using a test with $\alpha = 0.05/6$
2. Then we compare Medicare to Medicaid on the same outcome, also using $\alpha = 0.05/6$

. . . and then we do the other four comparisons, also at $\alpha = 0.05/6$.

Then across these six comparisons, our overall α can be (at worst)

- $0.05/6 + 0.05/6 + 0.05/6 + 0.05/6 + 0.05/6 + 0.05/6 = 0.05$
- So by changing our nominal confidence level from 95% to 99.167% in each comparison, we wind up with at least 95% confidence across this set of simultaneous comparisons.
- This is a conservative (worst case) approach.

Bonferroni approach for Pairwise Comparisons

Goal: Simultaneous p values comparing each pair of insurance types:

- Medicare vs Commercial
- Medicare vs Medicaid
- Medicare vs Uninsured
- Commercial vs Medicaid
- Commercial vs Uninsured
- Medicaid vs Uninsured

Bonferroni results for dbp by insurance

```
pairwise.t.test(dm192$dbp, dm192$insurance,
                p.adjust="bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  dm192$dbp and dm192$insurance

           commercial medicaid medicare
medicaid   0.31337    -         -
medicare   1.00000    0.00082   -
uninsured  1.00000    1.00000   0.91293

P value adjustment method: bonferroni
```

Tukey's Honestly Significant Differences

Most appropriate for **pre-planned** comparisons, with a balanced (or near-balanced) design.

Goal: Simultaneous (less conservative) confidence intervals and p values for our six pairwise comparisons:

- Medicare vs Commercial
- Medicare vs Medicaid
- Medicare vs Uninsured
- Commercial vs Medicaid
- Commercial vs Uninsured
- Medicaid vs Uninsured

Tukey HSD Confidence Intervals

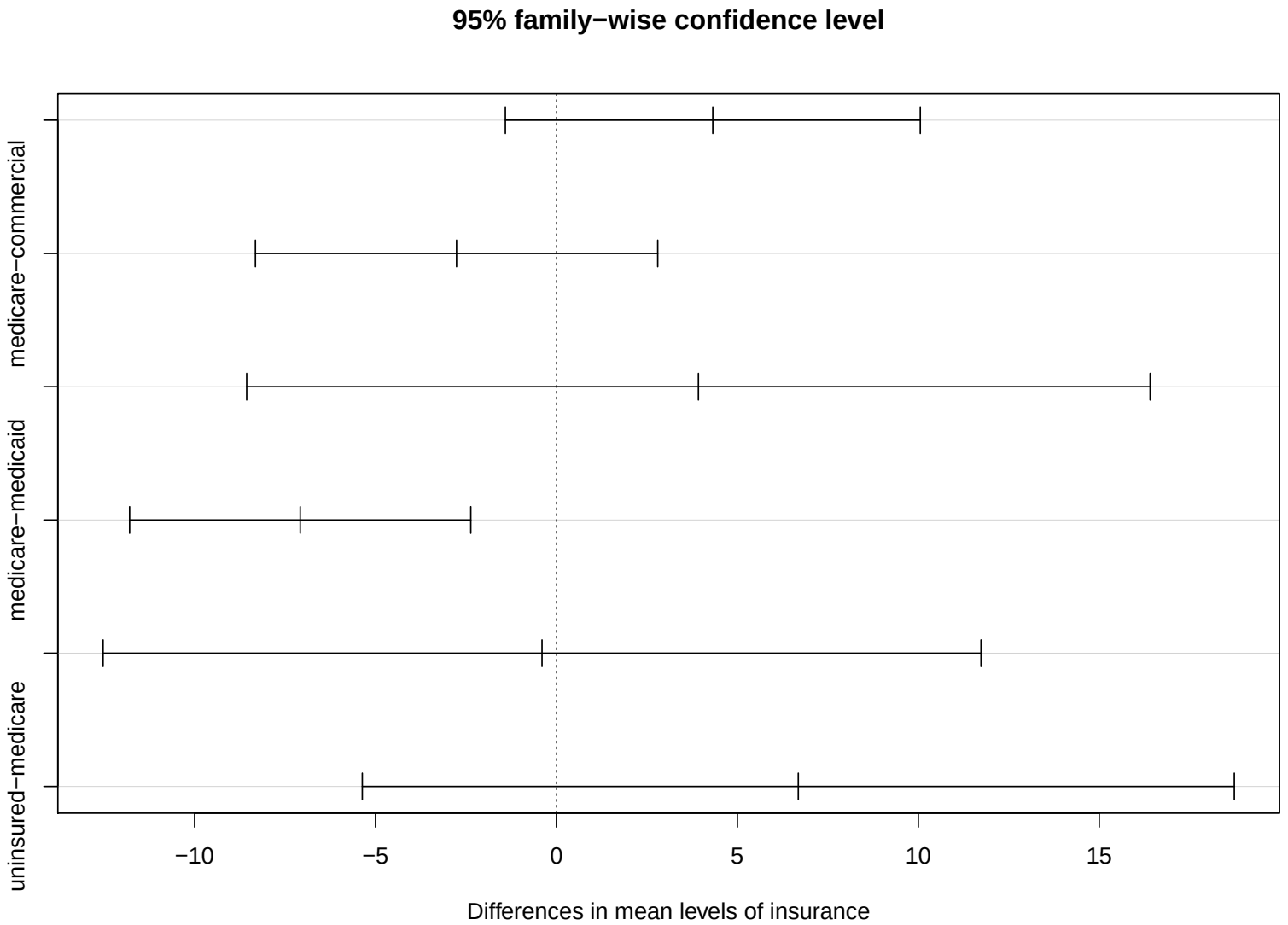
```
TukeyHSD(aov(dbp ~ insurance, data = dm192))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = dbp ~ insurance, data = dm192)

$insurance
                          diff        lwr       upr     p adj
medicaid-commercial   4.321087  -1.412308 10.054482 0.2095130
medicare-commercial  -2.760256  -8.319617  2.799104 0.5723295
uninsured-commercial  3.923077  -8.560153 16.406307 0.8475052
medicare-medicaid    -7.081343 -11.795510 -2.367177 0.0007847
uninsured-medicaid   -0.398010 -12.528463 11.732443 0.9997788
uninsured-medicare    6.683333  -5.365840 18.732506 0.4774431
```

Plot of Tukey HSD results (default, no relabeling)



Need to build smaller names for `insurance` levels

The `forcats` package can help

```
levels(dm192$insurance)
```

```
[1] "commercial" "medicaid"   "medicare"   "uninsured"
```

```
dm192$ins <- fct_recode(dm192$insurance,
                        "C" = "commercial",
                        "Md" = "medicaid",
                        "Mr" = "medicare",
                        "U" = "uninsured")
levels(dm192$ins)
```

```
[1] "C"  "Md" "Mr" "U"
```

Tukey 90% HSD CI

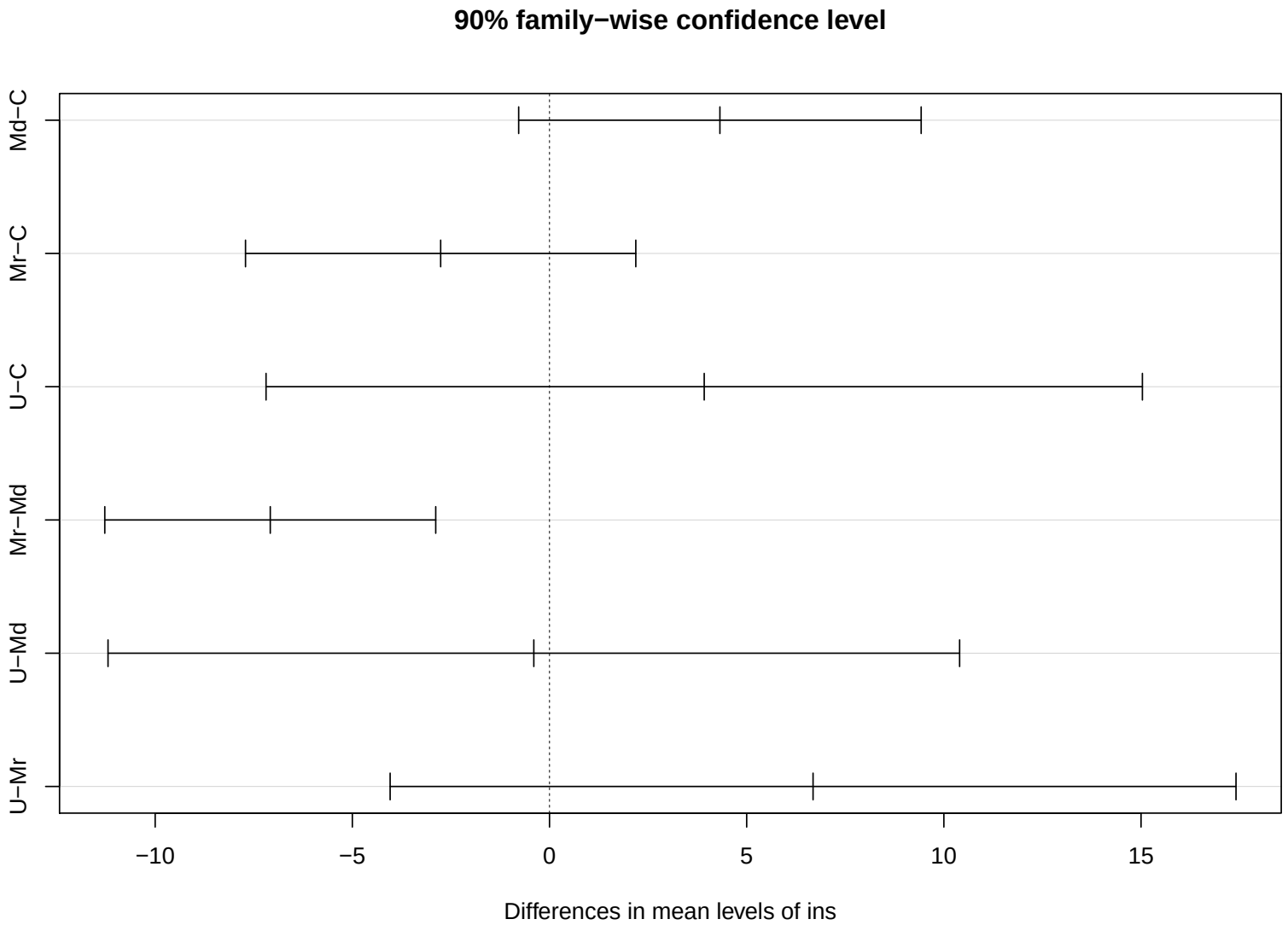
```
TukeyHSD(aov(dbp ~ ins, data = dm192), conf.level = 0.9)
```

```
  Tukey multiple comparisons of means
    90% family-wise confidence level

Fit: aov(formula = dbp ~ ins, data = dm192)

$ins
             diff          lwr       upr     p adj
Md-C    4.321087  -0.7822561  9.424430 0.2095130
Mr-C   -2.760256  -7.7086896  2.188177 0.5723295
U-C     3.923077  -7.1883505 15.034504 0.8475052
Mr-Md -7.081343 -11.2774623 -2.885224 0.0007847
U-Md   -0.398010 -11.1954283 10.399408 0.9997788
U-Mr    6.683333  -4.0417366 17.408403 0.4774431
```

Plot of 90% Tukey HSD Intervals



Conclusions for dbp by insurance

The dbp levels are statistically significantly higher in some insurance groups than in others.

In particular, with 90% confidence across all six pairwise comparisons of insurance types, we see a statistically significant difference between Medicare and Medicaid, with Medicare patients showing dbp levels that are 7.1 mm Hg lower on average than Medicaid patients (90% simultaneous CI: 2.9 to 11.3 mm Hg.)

Looking at dbp by practice

```
summary(aov(dbp ~ practice, data = dm192))
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
practice      3   2694   898.0   7.714 6.9e-05 ***
Residuals   188  21887   116.4
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bonferroni *p* values for dbp by practice

```
pairwise.t.test(dm192$dbp, dm192$practice,
                p.adjust="bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  dm192$dbp and dm192$practice

  A      B      C
B 1.0000 -      -
C 0.0044 0.0014 -
D 0.0174 0.0063 1.0000

P value adjustment method: bonferroni
```

Tukey HSD CI for dbp by practice

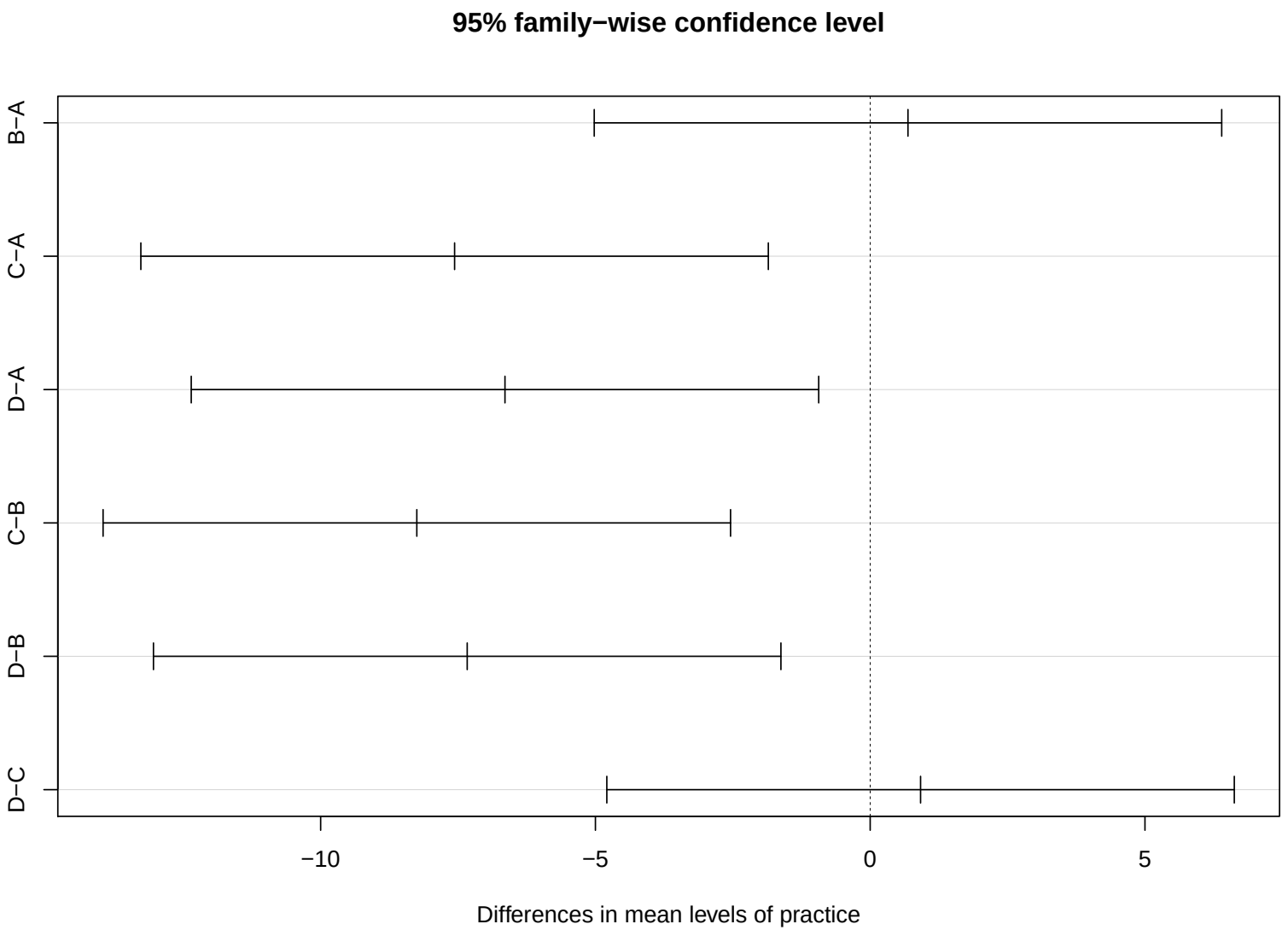
```
TukeyHSD(aov(dbp ~ practice, data = dm192))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = dbp ~ practice, data = dm192)

$practice
           diff         lwr        upr     p adj
B-A   0.6875000  -5.021573  6.3965730 0.9894298
C-A -7.5625000 -13.271573 -1.8534270 0.0040503
D-A -6.6458333 -12.354906 -0.9367604 0.0152566
C-B -8.2500000 -13.959073 -2.5409270 0.0013559
D-B -7.3333333 -13.042406 -1.6242604 0.0057255
D-C  0.9166667  -4.792406  6.6257396 0.9756496
```

Plot of Tukey HSD Results (dbp by practice)



Conclusions for dbp by practice

The dbp levels are statistically significantly higher in some practices than in others.

In particular, with 95% confidence across all six pairwise comparisons of practices, we see statistically significant differences between A and C and between A and D, as well as between B and C and between B and D, with C and D showing significantly lower dbp than either A or B.

For example, comparing C to A, we see a difference of 7.6 mm Hg (with A higher than C), with 95% CI (via Tukey HSD) of (1.9, 13.3) mm Hg.

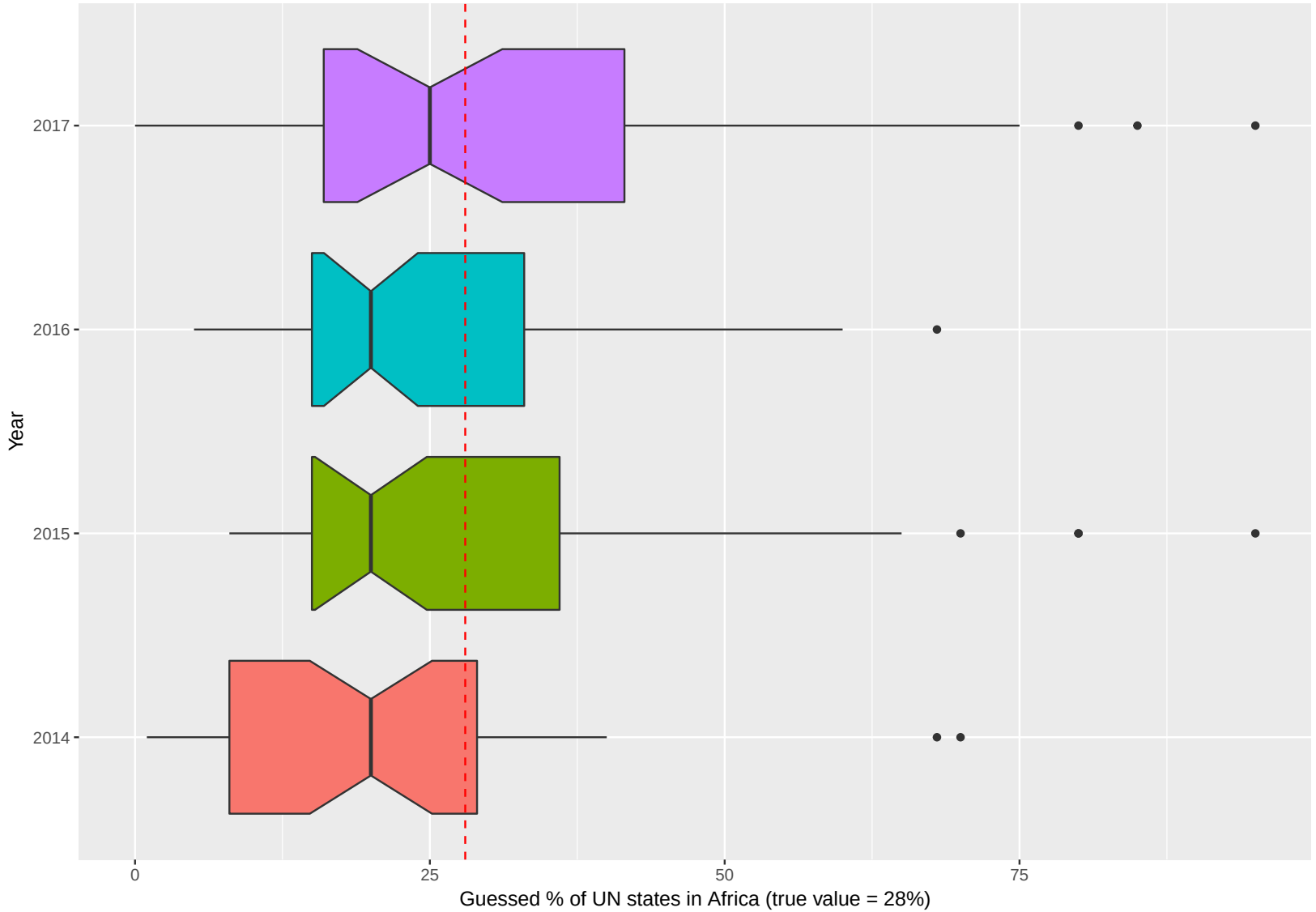
One Last Example

Here is the data from `class17a` again, Let's consider `africa.pct` by `year`

```
class17a %>%
  filter(!is.na(africa.pct)) %>%
  group_by(year) %>%
  summarise(n(), mean(africa.pct), sd(africa.pct))
```

```
# A tibble: 4 x 4
   year `n()` `mean(africa.pct)` `sd(africa.pct)`
  <int> <int>              <dbl>            <dbl>
1  2014    41            20.70732         15.58404
2  2015    49            28.79592         20.20518
3  2016    51            26.58824         15.96894
4  2017    43            31.09302         24.00081
```

Plot comparing four groups for `class17a`



ANOVA Question?

The question ANOVA answers is whether the means across the four subpopulations (here, years) are the same or not the same.

- Doesn't address the issue of which year best estimated the true value (28%) at all.

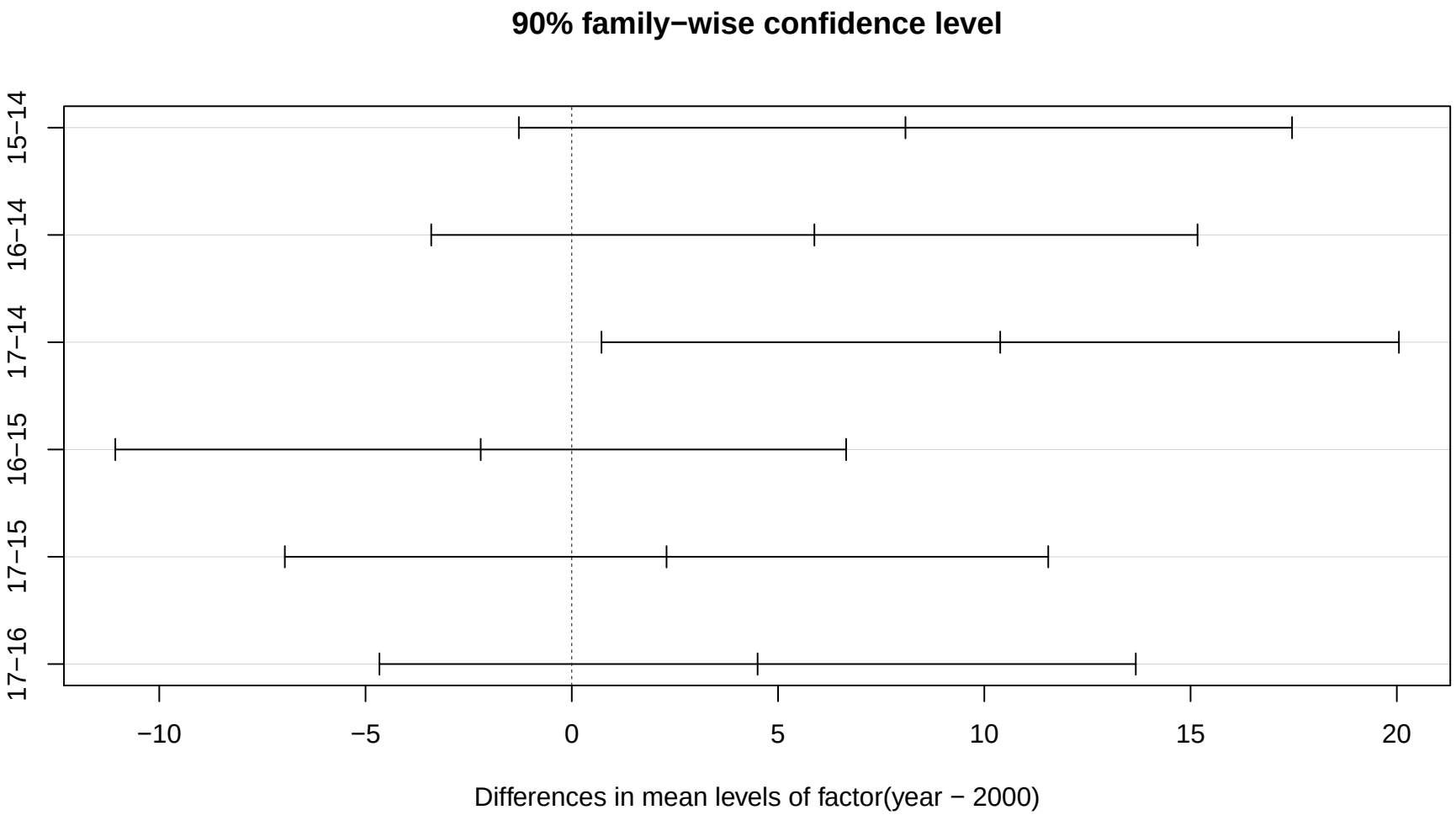
```
anova(lm(africa.pct ~ factor(year), data = class17a))
```

```
Analysis of Variance Table

Response: africa.pct
              Df Sum Sq Mean Sq F value  Pr(>F)
factor(year)   3   2509  836.45  2.2725 0.08174 .
Residuals    180  66254  368.08
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey HSD 90% Comparisons, plotted

```
plot(TukeyHSD(aov(africa.pct ~ factor(year-2000),
                  data = class17a), conf.level = 0.90))
```



ANOVA summary

If you want to compare more than two population means, and are willing to assume Normality, the Analysis of Variance is attractive.

- Equivalent to fitting a linear model with a categorical predictor
- Compare the means, overall, using an F test
- Assess individual pairwise comparisons with Bonferroni and Tukey HSD procedures
- If Normality is a serious issue, consider Kruskal-Wallis test (431) or bootstrap (432) approaches

On p values and statistical significance

The Value of a *p*-Valueless Paper

Jason T. Connor (2004) *American J of Gastroenterology* 99(9): 1638-40.

Abstract: As is common in current biomedical research, about 85% of original contributions in *The American Journal of Gastroenterology* in 2004 have reported *p*-values. However, none are reported in this issue's article by Abraham et al. who, instead, rely exclusively on effect size estimates and associated confidence intervals to summarize their findings. **Authors using confidence intervals communicate much more information in a clear and efficient manner than those using *p*-values. This strategy also prevents readers from drawing erroneous conclusions caused by common misunderstandings about *p*-values**. I outline how standard, two-sided confidence intervals can be used to measure whether two treatments differ or test whether they are clinically equivalent.

DOI: 10.1111/j.1572-0241.2004.40592.x

Editorial from JAMA Cardiology 2016-10-12

Editor's Note

Do Not Over (*P*) Value Your Research Article

Laine E. Thomas, PhD; Michael J. Pencina, PhD

***P* value** is by far the most prevalent statistic in the medical literature but also one attracting considerable controversy. Recently, the American Statistical Association[1] released a policy statement on *P* values, noting that misunderstanding and misuse of *P* values is an important contributing factor to the common problem of scientific conclusions that fail to be reproducible. Furthermore, reliance on *P* values may distract from the good scientific principles that are needed for high-quality research. Mark et al[2] delve deeper into the history and interpretation of the *P* value in this issue of *JAMA Cardiology*. Herein, we take the opportunity to state a few principles to help guide authors in the use and reporting of *P* values in the journal.

Related article

When the limitations surrounding *P* values are emphasized, a common question is, "What should we do instead?" Ron Wasserstein of the American Statistical Association explained: "In the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy.... Instead, journals [should evaluate] papers based on clear and detailed description of the study design, execution, and analysis, having conclusions that are based on valid statistical interpretations and scientific arguments, and reported transparently and thoroughly enough to be rigorously scrutinized by others."[3]

We suggest that researchers submitting manuscripts to *JAMA Cardiology* should also consider the following:

1. Data that are descriptive of the sample (ie, indicating imbalances between observed groups but not making inference to a population) should not be associated with *P* values. Appropriate language, in this case, would describe numerical differences and sample summary statistics and focus on differences of clinical importance.
2. In addition to summary statistics and confidence intervals, standardized differences (rather than *P* values) are a preferred way to exhibit imbalances between groups.
3. *P* values are most meaningful in the context of clear, a priori hypotheses that support the main conclusions of a manuscript.
4. Reporting stand-alone *P* values is discouraged, and preference should be given to presentation and interpretation of effect sizes and their uncertainty (confidence intervals) in the scientific context and in light of other evidence. Crossing a threshold (eg, $P < .05$) by itself constitutes only weak evidence.
5. Researchers should define and interpret effect measures that are clinically relevant. For example, clinical importance is often difficult to establish on the odds ratio scale but is clearer on the risk ratio or absolute risk difference scale.

In summary, following Mark et al,[2] we encourage researchers to focus on interpreting clinical research data in terms of treatment "effect" magnitude and precision, using *P* value only as one of many complementary tools in the statistical toolbox.

Mark, Lee, Harrell JAMA Cardiol 2016-10-12

Abstract

P values and hypothesis testing methods are frequently misused in clinical research. Much of this misuse appears to be owing to the widespread, mistaken belief that they provide simple, reliable, and objective triage tools for separating the true and important from the untrue or unimportant. The primary focus in interpreting therapeutic clinical research data should be on the treatment ("oomph") effect, a metaphorical force that moves patients given an effective treatment to a different clinical state relative to their control counterparts. This effect is assessed using 2 complementary types of statistical measures calculated from the data, namely, effect magnitude or size and precision of the effect size. In a randomized trial, effect size is often summarized using constructs, such as odds ratios, hazard ratios, relative risks, or adverse event rate differences. How large a treatment effect has to be to be consequential is a matter for clinical judgment. The precision of the effect size (conceptually related to the amount of spread in the data) is usually addressed with confidence intervals. *P* values (significance tests) were first proposed as an informal heuristic to help assess how "unexpected" the observed effect size was if the true state of nature was no effect or no difference. Hypothesis testing was a modification of the significance test approach that envisioned controlling the false-positive rate of study results over many (hypothetical) repetitions of the experiment of interest. Both can be helpful but, by themselves, provide only a tunnel vision perspective on study results that ignores the clinical effects the study was conducted to measure.

doi:10.1001/jamacardio.2016.3312

On Experiments

. . . the null hypothesis is never proved or established, but is possibly disapproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis.

- R. A. Fisher

Do not be too timid and squeamish about your actions. All life is an experiment. The more experiments, the better.

- Ralph Waldo Emerson

Why Dividing Data Comparisons into Categories based on Significance Levels is Terrible.

The common practice of dividing data comparisons into categories based on significance levels is terrible, but it happens all the time. . . . so it's worth examining the prevalence of this error.

Link to Andrew Gelman's blog, 2016-10-15

Gelman on *p* values, 1

Let me first briefly explain why categorizing based on p-values is such a bad idea. Consider, for example, this division:

- "really significant" for $p < .01$,
- "significant" for $p < .05$,
- "marginally significant" for $p < .1$, and
- "not at all significant" otherwise.

Now consider some typical *p*-values in these ranges: say, $p = .005$, $p = .03$, $p = .08$, and $p = .2$.

Translate these two-sided *p*-values back into z-scores, which we can do in R via `qnorm(c(.005, .03, .08, .2)/2, lower.tail = FALSE)`

Gelman on *p* values, 2

Description	really sig.	sig.	marginally sig.	not at all sig.
p value	0.005	0.03	0.08	0.20
Z score	2.8	2.2	1.8	1.3

The seemingly yawning gap in p-values comparing the "not at all significant" p-value of .2 to the "really significant" p-value of .005, is only 1.5.

If you had two independent experiments with z-scores of 2.8 and 1.3 and with equal standard errors and you wanted to compare them, you'd get a difference of 1.5 with a standard error of 1.4, which is completely consistent with noise.

Gelman on *p* values, 3

From a **statistical** point of view, the trouble with using the p-value as a data summary is that the p-value is only interpretable in the context of the null hypothesis of zero effect — and (much of the time), nobody's interested in the null hypothesis.

Indeed, once you see comparisons between large, marginal, and small effects, the null hypothesis is irrelevant, as you want to be comparing effect sizes.

From a **psychological** point of view, the trouble with using the p-value as a data summary is that this is a kind of deterministic thinking, an attempt to convert real uncertainty into firm statements that are just not possible (or, as we would say now, just not replicable).

The key point: The difference between statistically significant and NOT statistically significant is not, generally, statistically significant.