Stat 411/511

Comparison of Many Groups

Nov 2 2015

Charlotte Wickham

stat511.cwick.co.nz

More groups

So far we have only looked at **two groups**, now we'll look at **multiple groups**.

In this chapter two big questions:

1. How can we compare two groups in particular?

   back to the *t*-test

2. Are there **any** differences in the population means?

   Extra Sum of Squares F-test (one way ANOVA)

Next chapter:

3. How different are the means?

   Linear combinations and multiple comparisons

   Chapter 6

The “groups” could arise:

as samples from different populations

we are interested in making inference about the means of the populations we sampled from

as observations from different treatments assigned to units at random

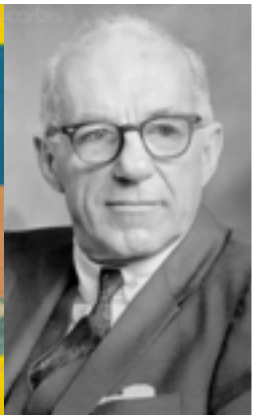
we are interested in making inference about the additive treatment effects of the treatments

we could consider randomization procedures, but like in the two group case, sampling models give good approximations

translate differences in population means to differences in additive treatment effects

Case 0502

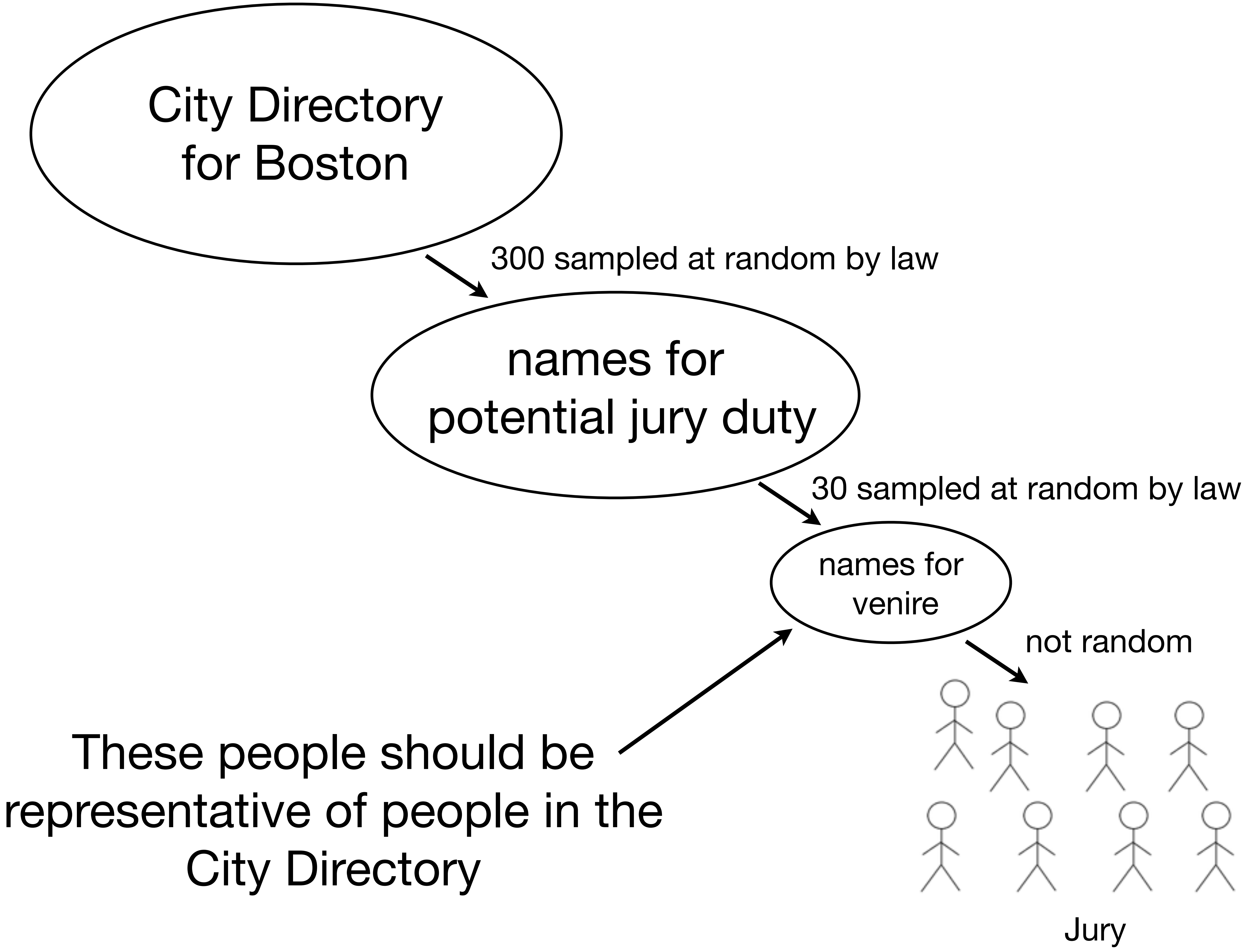
Spock Trial case study



Dr. Benjamin Spock was tried on charges of conspiring to violate the Selective Service Act.

There were no women on his jury.

Did the judge in the trial have *venires* that systematically underrepresented women?



City Directory for Boston

300 sampled at random by law

names for potential jury duty

30 sampled at random by law

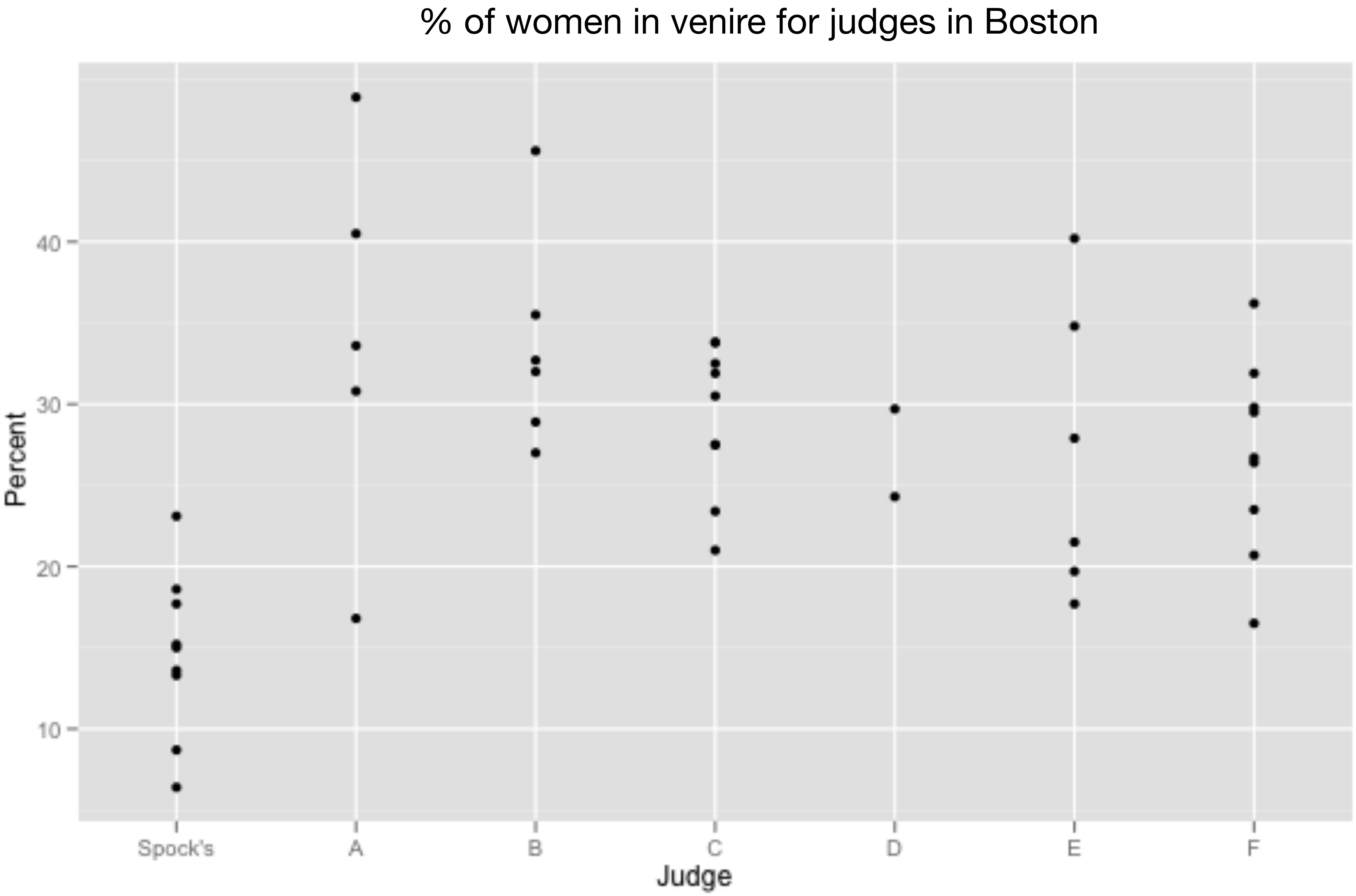
names for venire

not random

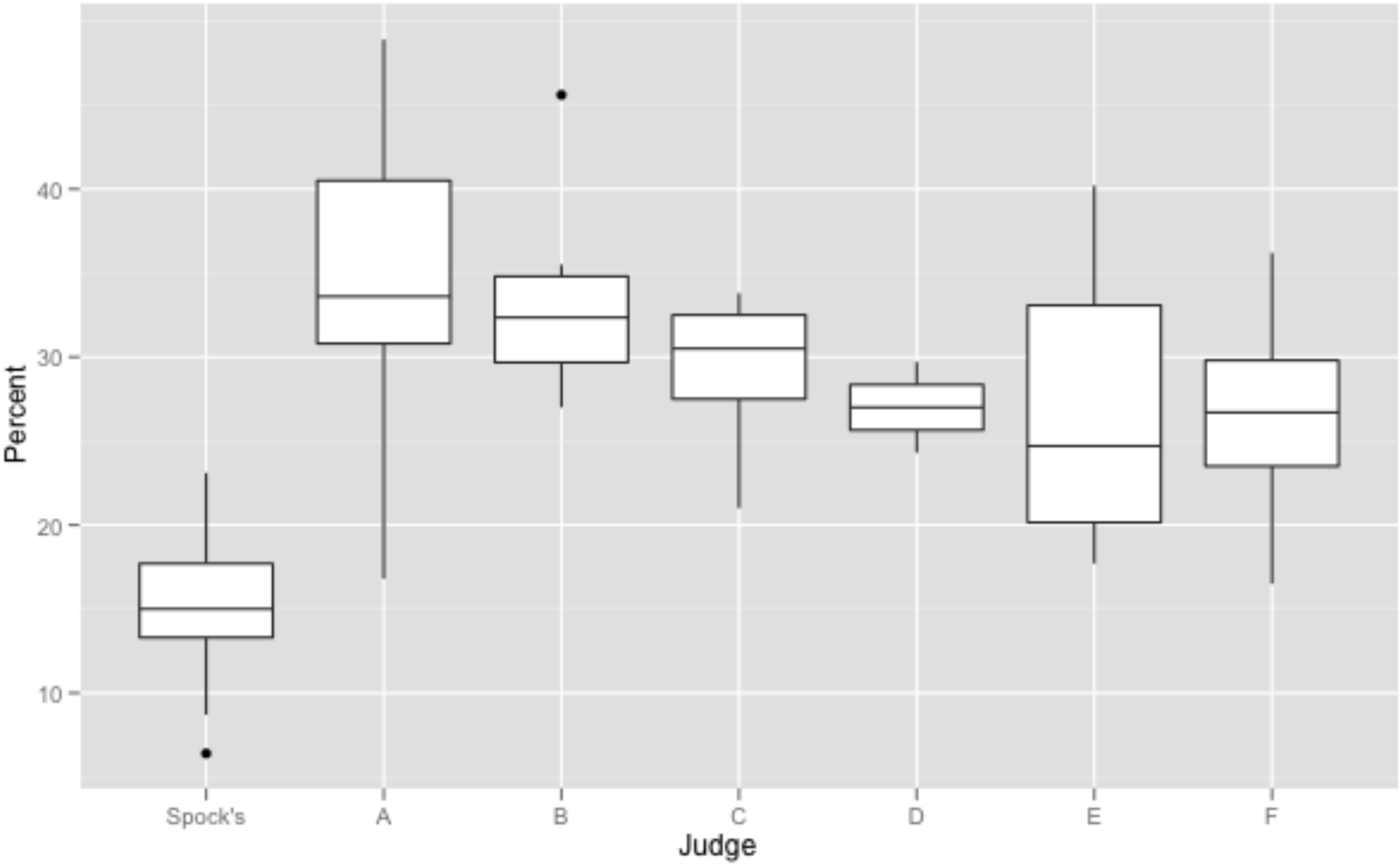
Jury

These people should be representative of people in the City Directory

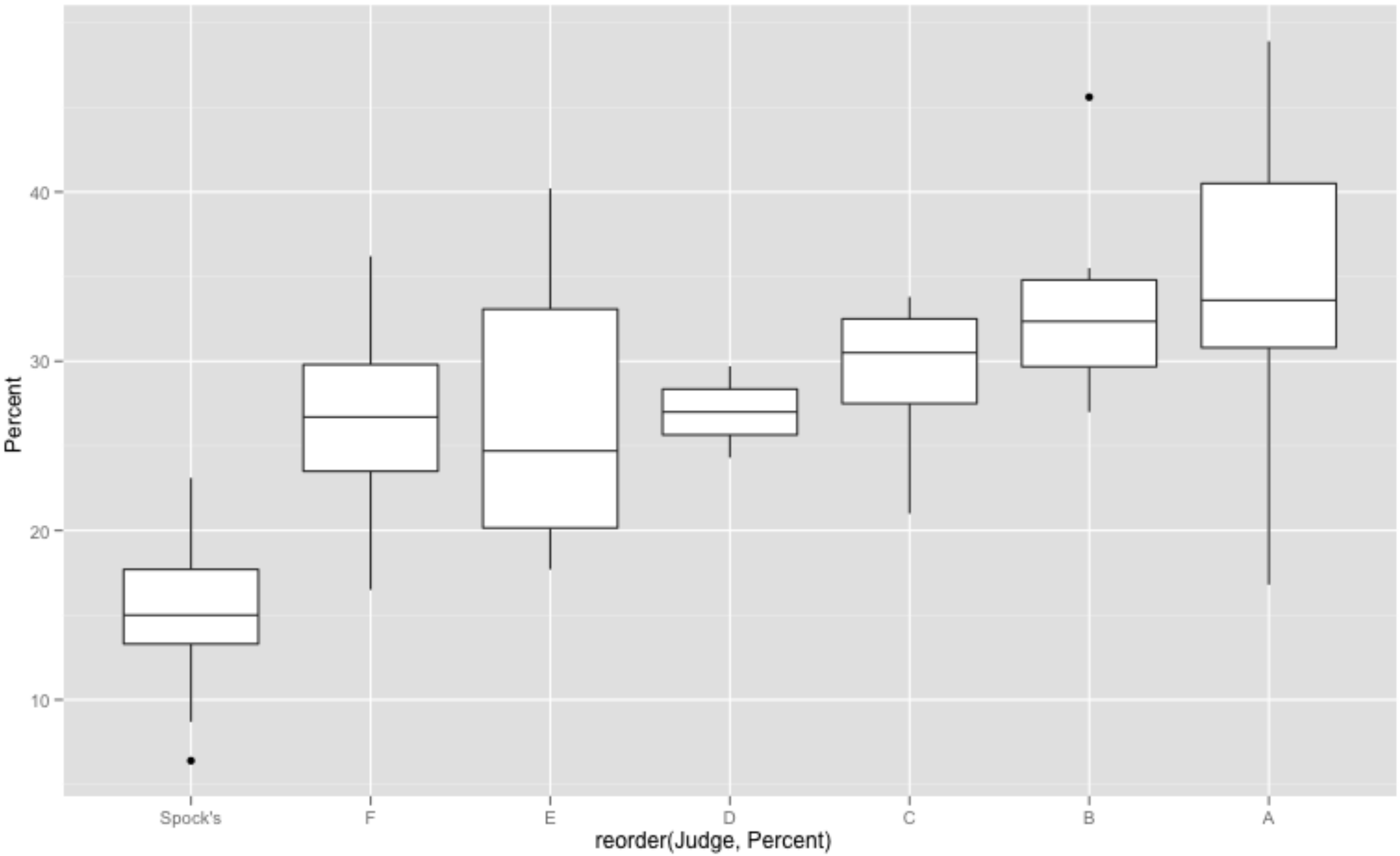
```
library(Sleuth2)
library(ggplot2)
qplot(Judge, Percent, data = case0502)
```



```
qplot(Judge, Percent, data = case0502, geom = "boxplot")
```



```
qplot(reorder(Judge, Percent), Percent, data = case0502, geom = "boxplot")
```



```
qplot(Percent, data = case0502)
      + facet_wrap(~ Judge, ncol = 1)
```

Some notation

I = number of groups

μ_1 = population mean of population 1.

σ_1 = population standard deviation of population 1.

$\bar{Y}_1$ = sample average of group 1.

s_1 = sample standard deviation of group 1.

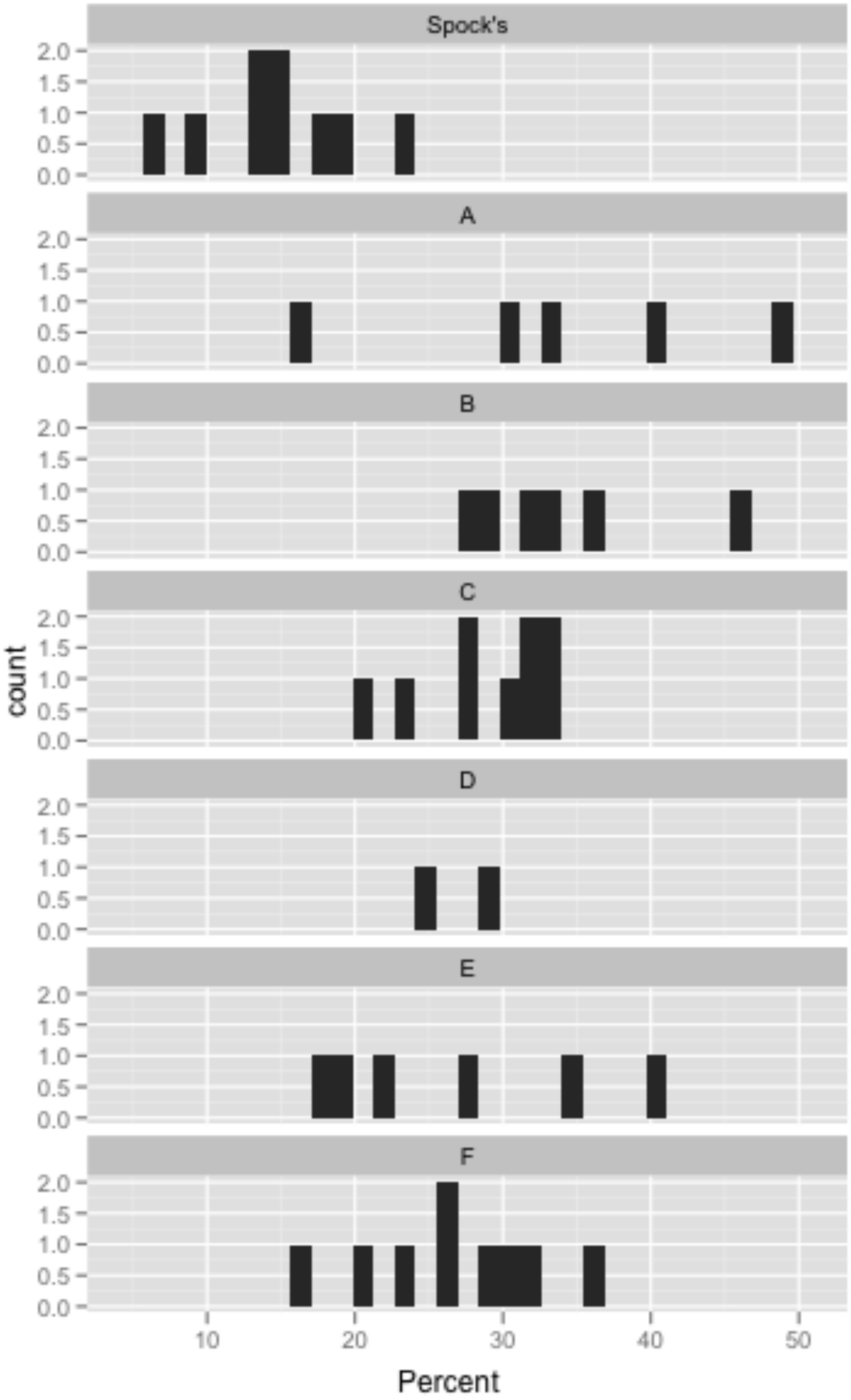
and so on...

μ_I = population mean of population I.

σ_I = population standard deviation of population I.

$\bar{\bar{Y}}_I$ = sample average of group I.

s_I = sample standard deviation of group I.



Assumptions

of ANOVA F-test,
two group t comparisons &
linear combinations

1. Normally distributed populations.
2. Equal population standard deviations, $\sigma_1 = \sigma_2 = \ldots = \sigma_I = \sigma$.
3. Independence of subjects between and within groups.

Same as the two-sample t-test,
but now many samples.

in lab

Shortcut

sample averages

```
> with(case0502, tapply(Percent, Judge, mean))
       A        B        C        D        E        F  Spock's
34.12000 33.61667 29.10000 27.00000 26.96667 26.80000 14.62222
```

sample sds

```
> with(case0502, tapply(Percent, Judge, sd))
        A         B         C         D         E         F   Spock's
11.941817  6.582224  4.592929  3.818377  9.010142  5.968878  5.038794
```

sample sizes

```
> with(case0502, tapply(Percent, Judge, length))
      A       B       C       D       E       F Spock's
      5       6       9       2       6       9       9
```

sample sizes,

not counting missing values

```
> with(case0502, tapply(Percent, Judge,
                                function(x) sum(!is.na(x))))
```

```
averages <- with(case0502, tapply(Percent, Judge, mean))
sds <- with(case0502, tapply(Percent, Judge, sd))
ns <- with(case0502, tapply(Percent, Judge, length))
```

save for later...

Your turn

```
> head(case0501)
  Lifetime Diet
1     35.5   NP
2     35.4   NP
3     34.9   NP
4     34.8   NP
5     33.8   NP
6     33.5   NP
```

How would you get a sample average for each Diet group?

How would you get a sample median for each Diet group?

From previous slide:

```
> with(case0502, tapply(Percent, Judge, mean))
```

Comparing two groups

Comparing two groups

Did Spock's judge have less women in his venire population than judge A?

Null: $\mu_{\text{spock}} = \mu_{\text{A}}$,

μ = mean percent of women.

Two sample *t*-statistic: $$\frac{\overline{Y}_2 - \overline{Y}_1}{SE_{\overline{Y}_2 - \overline{Y}_1}}$$ same as before

$$\mathrm{SE}_{\overline{Y}_2 - \overline{Y}_1} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$ same as before

↗ involves the pooled standard deviation

Pooled standard deviation

Use **all** the groups (even if you only want to compare two!)

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \ldots (n_I - 1)}}$$

different

$$= \sqrt{\sum_{i=1}^{I}(n_i - 1)s_i^2 / \sum_{i=1}^{I}(n_i - 1)}$$

```
# pooled standard deviation
sp <- sqrt(sum((ns - 1) * sds^2) / sum(ns - 1))
```

different

degrees of freedom =
total number of subjects - number of groups =
n - I

We assume all groups come from **populations with the same standard deviation**, σ.

Each sample provides an estimate of σ with it’s sample standard deviation.

Even if we aren’t interested in the means of some groups, they still give us information about the population standard deviation.

We get the **best estimate** of, σ, by making use of **all the information** we have available, and pooling all the sample standard deviations.

This wouldn’t be a good idea if the other samples look like they come from populations with different standard deviations.

Did Spock's judge have less women in his venire than judge A?

Two sample *t*-statistic: $\frac{\overline{Y}_{spock} - \overline{Y}_A}{SE_{\overline{Y}_{spock} - \overline{Y}_A}}$

```
t.stat <- (averages[7] - averages[1] ) / (sp * sqrt(1/ns[7] + 1/ns[1]))

-5.055741
```

↑

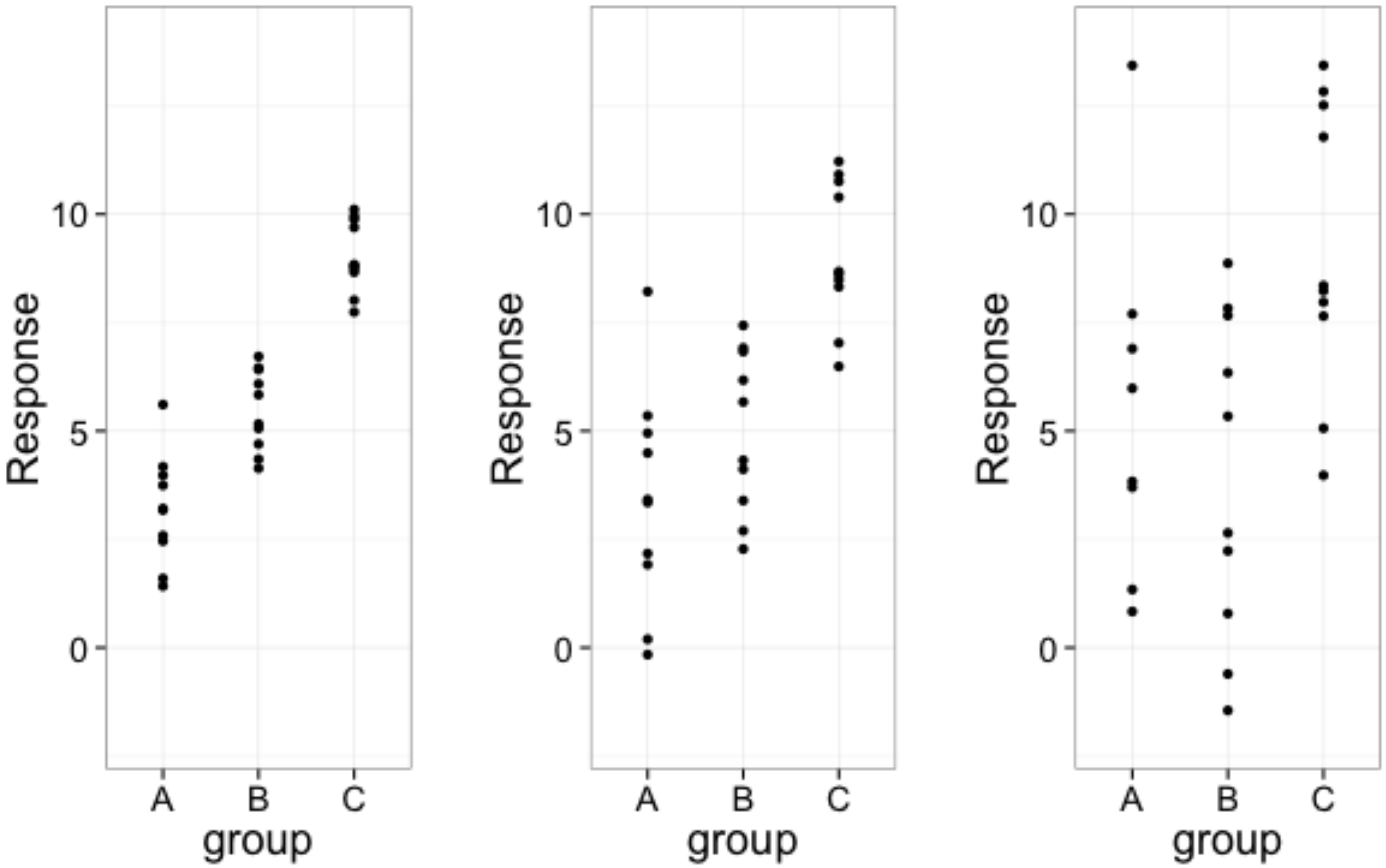
[subsetting see R code for safer way

Under the **null**: Has a Student's t-distribution with n-I degrees of freedom.

```
2*pt(t.stat, sum(ns) - length(ns))

 5.25123e-06
```

p-value



Consider these three datasets.

Which would you say gives more evidence of the groups coming from populations with different means?