Analysis of Variance

(ANOVA)

Wolf River

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including hexachlorobenzene (HCB).
- HCB known to cause various cancers and birth defects.

Wolf River study

- Standard method to test whether HCB is present in a river is to take samples at middepth.
- HCB is denser than water, so is it found at different concentrations at different depths?

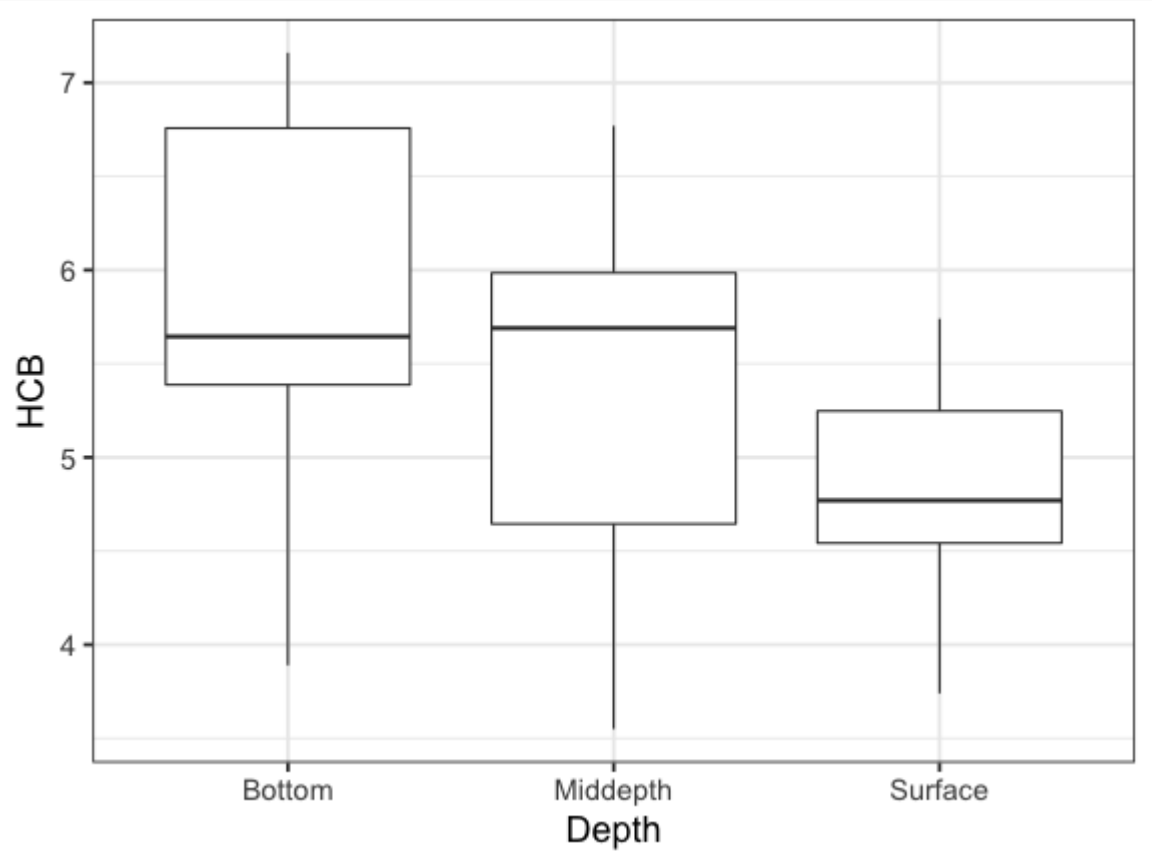
```
str(wolf)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ Depth : Factor w/ 3 levels "Bottom","Middepth",..:
##  $ Aldrin: num  3.08 3.58 3.81 4.31 4.35 4.4 3.67 5.17
##  $ HCB   : num  3.74 4.61 4 4.67 4.87 5.12 4.52 5.29 5
```

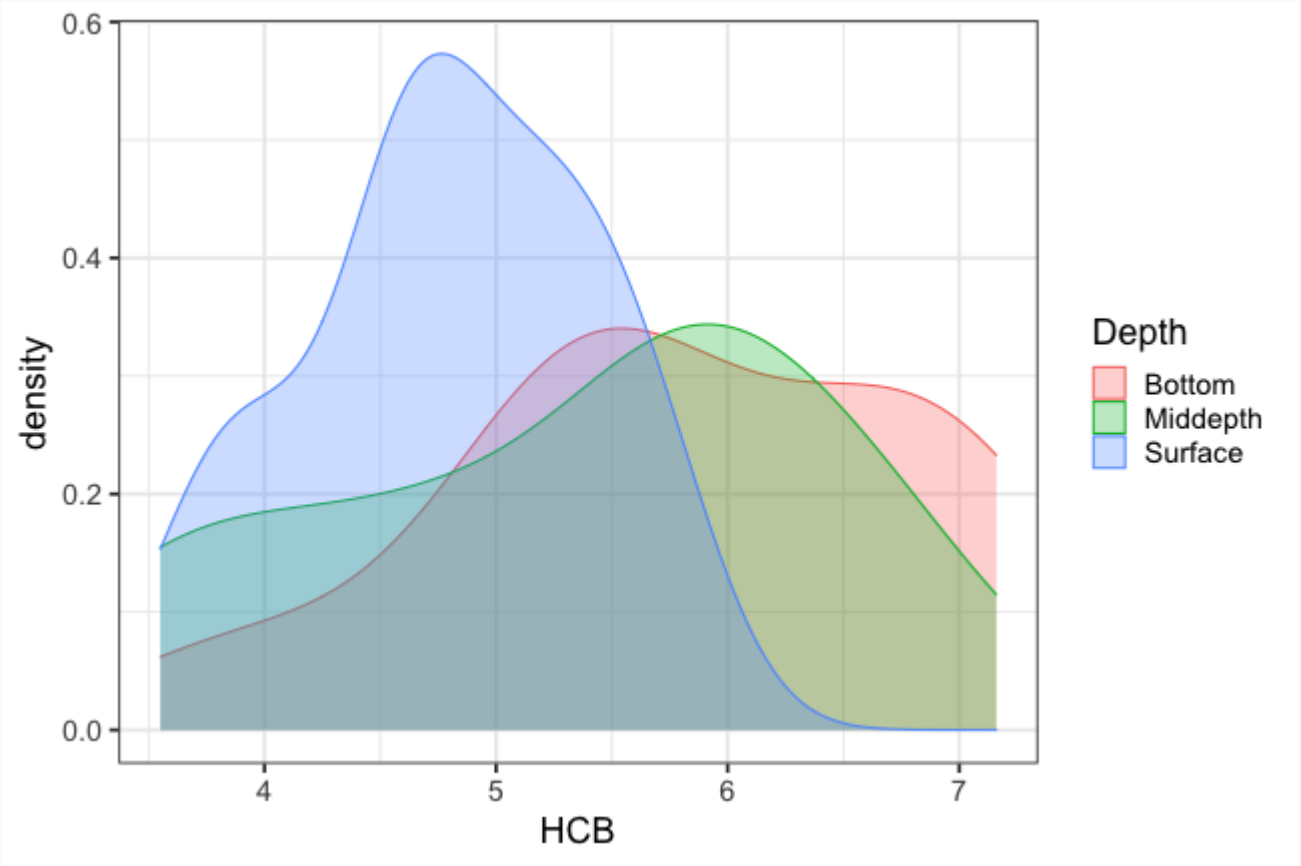
```
head(wolf)
```

```
##     Depth Aldrin  HCB
## 1 Surface   3.08 3.74
```

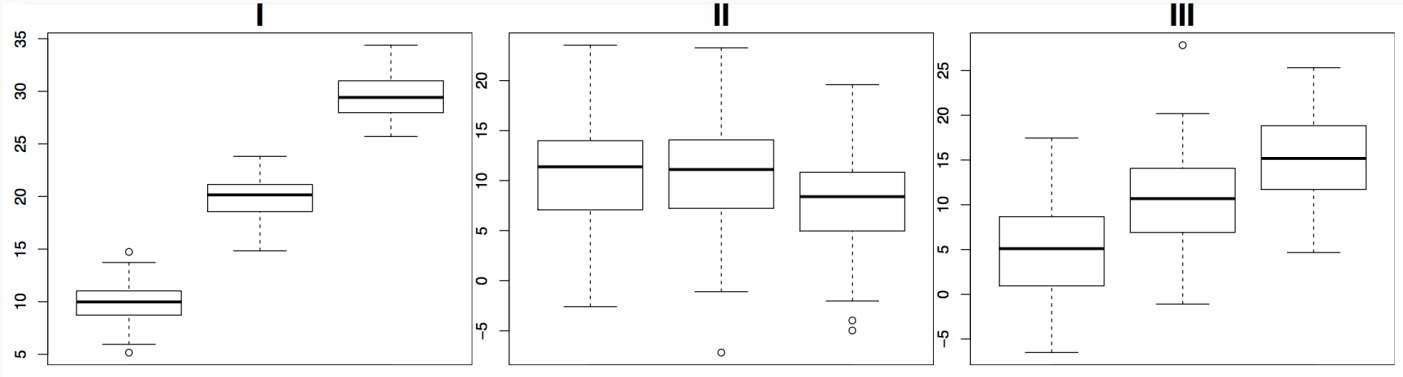
Wolf River data



Wolf River data, cont.



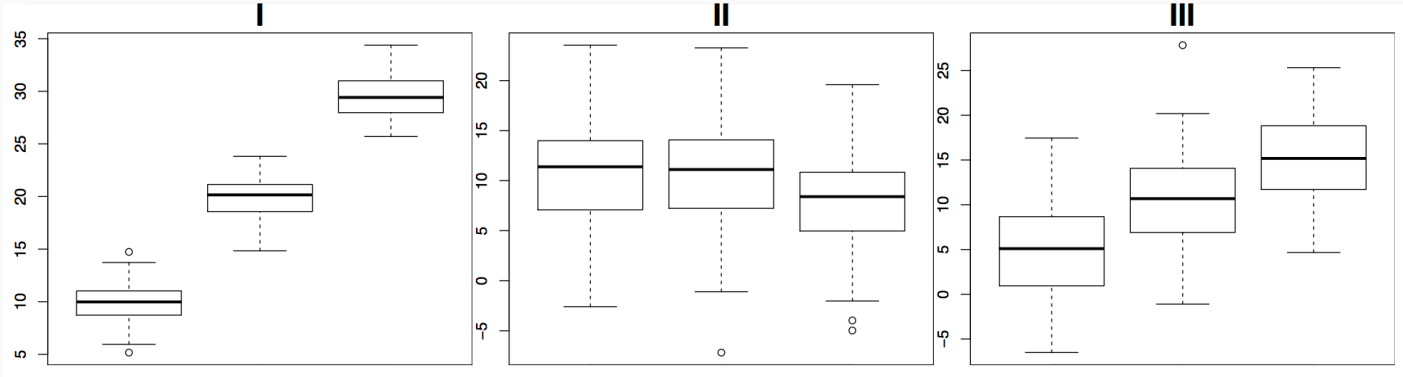
Which of the following plots shows groups with means that are *most* and *least* likely to be significantly different from each other?



1. most: I, least: II
2. most: II, least: III
3. most: I, least: III
4. most: III, least: II

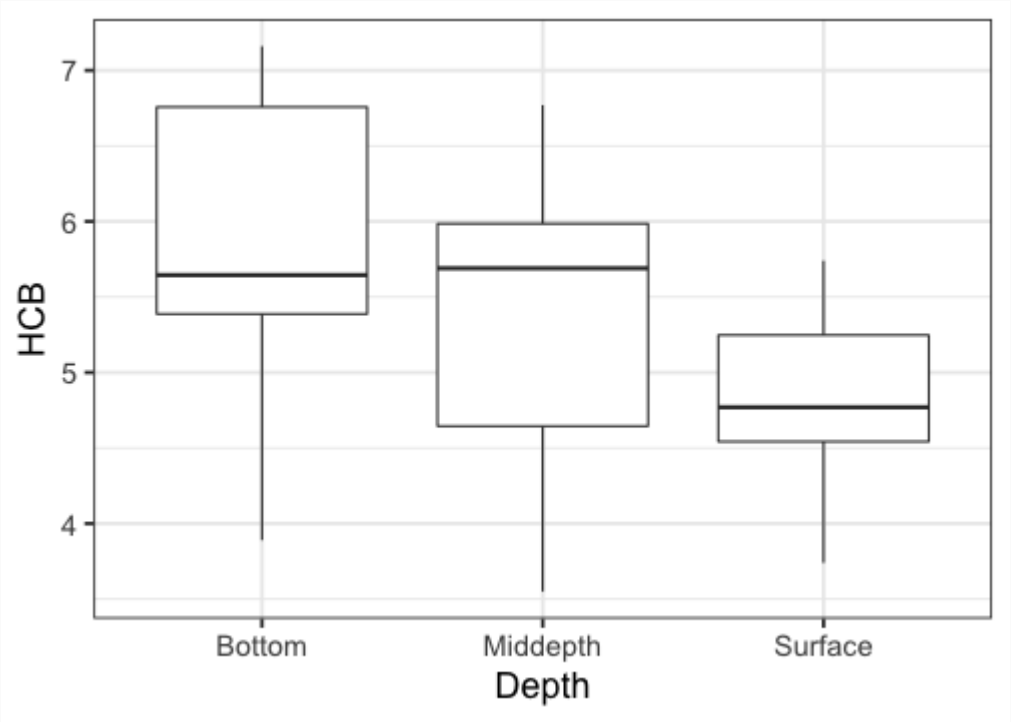
Constructing a statistic

Which of the following plots shows groups with means that are *most* and *least* likely to be significantly different from each other?



- I has a high F.
- II has a low F.
- III has a middling F.

Wolf River data




```
##             Df Sum Sq Mean Sq F value
## Depth        2   5.36   2.678    3.03
## Residuals   27  23.85   0.883
```

How big is 3.032?

ANOVA F-test

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_A : \text{At least one } \mu_j \text{ is different}$$

We can find the distribution of the F-statistic under the null hypothesis by

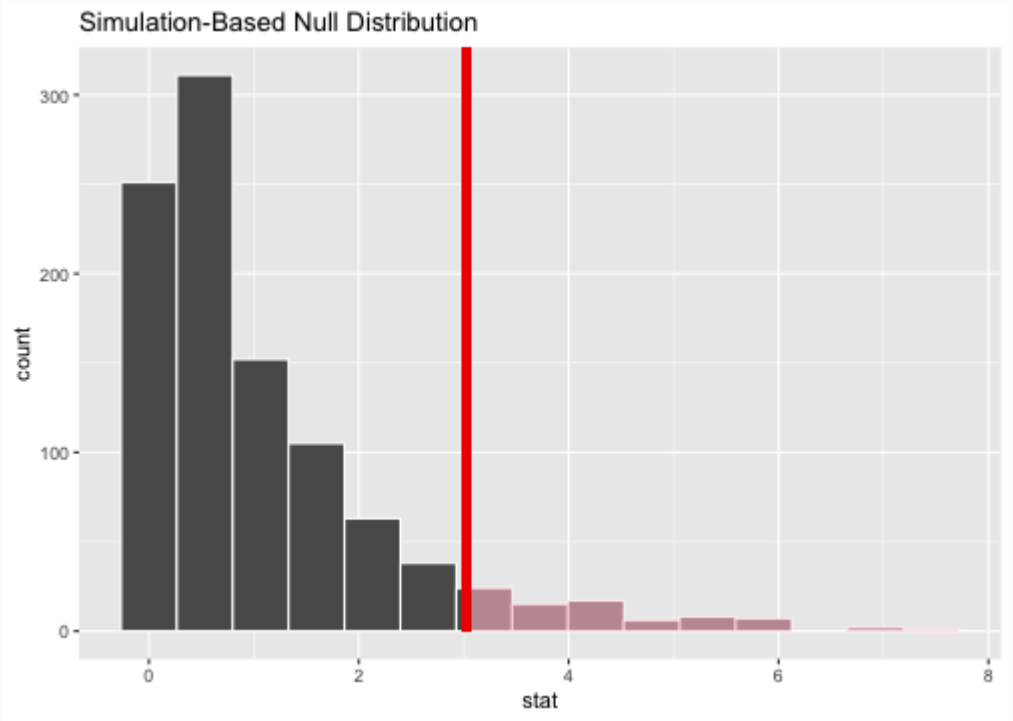
- Permutation
- Mathematical approximation

Sampling dist for F via Randomization

```
null <- wolf %>%
  specify(response = HCB, explanatory = Depth) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "F")
```

Sampling dist for F via Randomization, cont.

```
null %>%
  visualize() +
  shade_p_value(obs_stat = f$F[1],
                direction = "right")
```



Sampling dist for F via Randomization, cont.

```
null %>%
  get_p_value(obs_stat = f$F[1],
              direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.074
```

Sampling dist for F via Approximation

If:

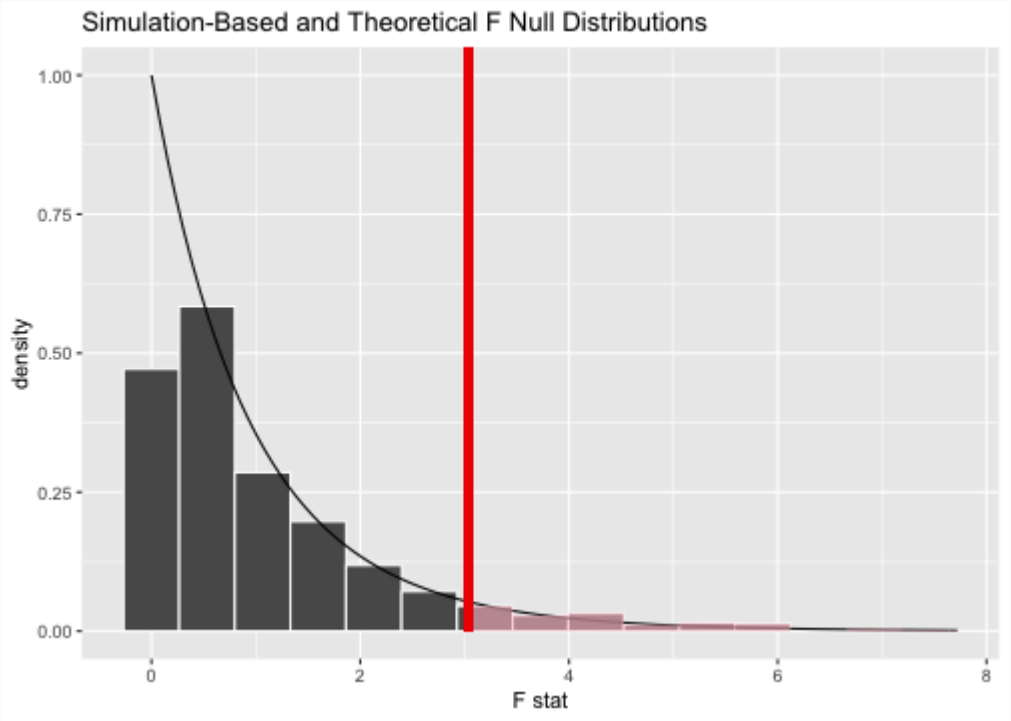
1. Independent observations.
2. Approximate normal distributions within groups.
3. Constant variance between groups.

Then the sampling distribution for the F statistic under the H_0 is well approximated by an F distribution with $df_1 = k - 1$ and $df_2 = N - k$. The p-value is represented by the upper tail.

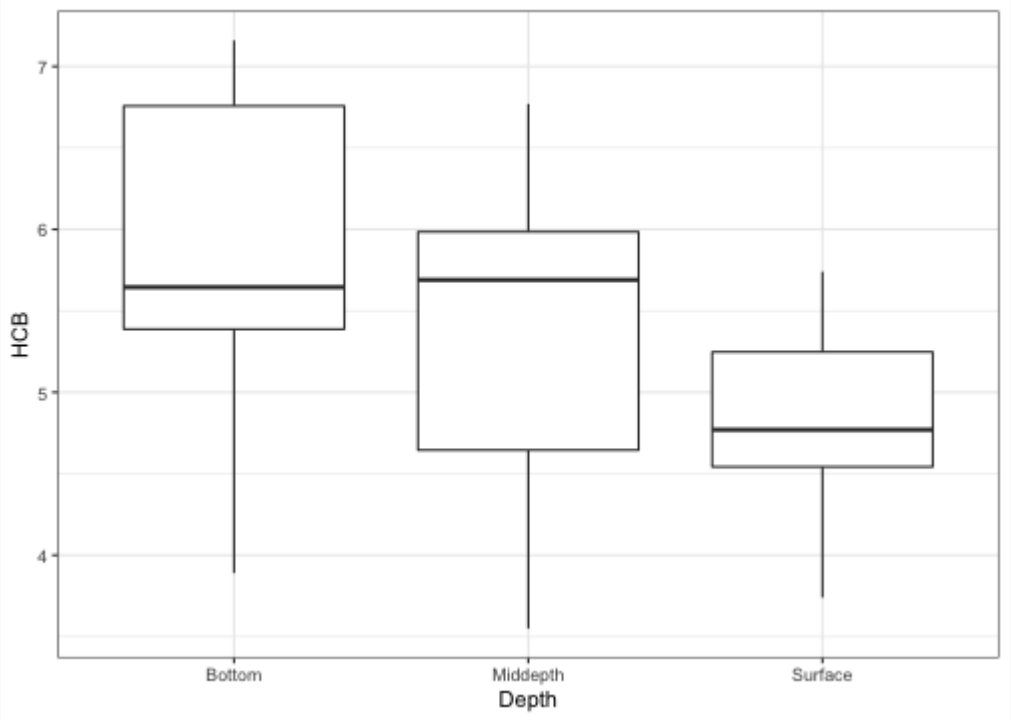
```
1 - pf(f$F[1], df1 = 2, df2 = 30 - 3)
```

```
## [1] 0.0649
```

```
null %>%
  visualize(method = "both") +
  shade_p_value(obs_stat = f$F[1],
                direction = "right")
```



Wolf River Conclusions



- With a p-value of ≈ 0.07, it is questionable whether HCB concentration functions the same at all three depths.
- *Replicating the study* could add some certainty.
- In a subsequent study, we may wish to only test middepth versus bottom.