Test statistics and randomization distributions

Applied Statistics and Experimental Design

Chapter 2

Peter Hoff

Statistics, Biostatistics and the CSSS
University of Washington

Example: wheat yield

Question: Is one fertilizer better than another, in terms of yield?

Outcome variable: Wheat yield.

Factor of interest: Fertilizer type, A or B. One **factor** having two **levels**.

Experimental material: One plot of land, divided into 2 rows of 6 subplots each.

Example: wheat yield

Question: Is one fertilizer better than another, in terms of yield?

Outcome variable: Wheat yield.

Factor of interest: Fertilizer type, A or B. One **factor** having two **levels**.

Experimental material: One plot of land, divided into 2 rows of 6 subplots each.

Example: wheat yield

Question: Is one fertilizer better than another, in terms of yield?

Outcome variable: Wheat yield.

Factor of interest: Fertilizer type, A or B. One **factor** having two **levels**.

Experimental material: One plot of land, divided into 2 rows of 6 subplots each.

Example: wheat yield

Question: Is one fertilizer better than another, in terms of yield?

Outcome variable: Wheat yield.

Factor of interest: Fertilizer type, A or B. One **factor** having two **levels**.

Experimental material: One plot of land, divided into 2 rows of 6 subplots each.

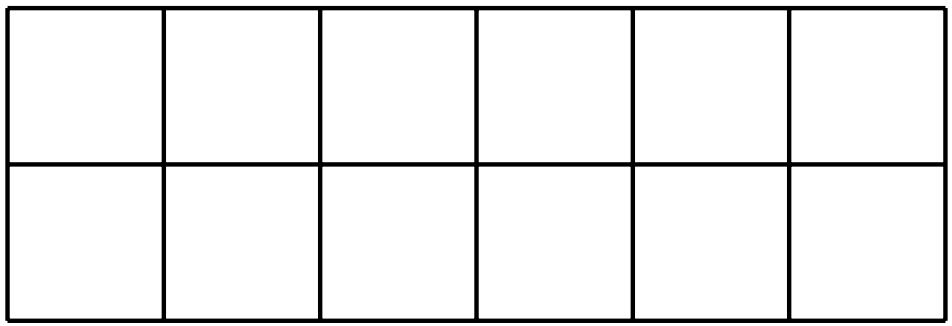
Example: wheat yield

Question: Is one fertilizer better than another, in terms of yield?

Outcome variable: Wheat yield.

Factor of interest: Fertilizer type, A or B. One **factor** having two **levels**.

Experimental material: One plot of land, divided into 2 rows of 6 subplots each.



Design questions

How should we assign treatments/factor levels to the plots?

Want to avoid confounding treatment effect with another source of variation.

Potential sources of variation: Fertilizer , soil , sun , water, etc.

Design questions

How should we assign treatments/factor levels to the plots?

Want to avoid confounding treatment effect with another source of variation.

Potential sources of variation: Fertilizer , soil , sun , water, etc.

Design questions

How should we assign treatments/factor levels to the plots?

Want to avoid confounding treatment effect with another source of variation.

Potential sources of variation: Fertilizer , soil , sun , water, etc.

Design questions

How should we assign treatments/factor levels to the plots?

Want to avoid confounding treatment effect with another source of variation.

Potential sources of variation: Fertilizer , soil , sun , water, etc.

Implementation of experiment

Assigning treatments *randomly* avoids any pre-experimental bias in results.

12 playing cards, 6 red, 6 black were shuffled and dealt:

$$\begin{aligned} \text{1st card black} &\longrightarrow \text{1st plot gets } B \\ \text{2nd card red} &\longrightarrow \text{2nd plot gets } A \\ \text{3rd card black} &\longrightarrow \text{3rd plot gets } B \\ &\vdots \end{aligned}$$

This is the first design we will study, a **completely randomized design**.

Implementation of experiment

Assigning treatments *randomly* avoids any pre-experimental bias in results.

12 playing cards, 6 red, 6 black were shuffled and dealt:

$$\begin{array}{rcl} \text{1st card black} & \longrightarrow & \text{1st plot gets } B \\ \text{2nd card red} & \longrightarrow & \text{2nd plot gets } A \\ \text{3rd card black} & \longrightarrow & \text{3rd plot gets } B \\ & \vdots & \end{array}$$

This is the first design we will study, a **completely randomized design**.

Implementation of experiment

Assigning treatments *randomly* avoids any pre-experimental bias in results.

12 playing cards, 6 red, 6 black were shuffled and dealt:

$$\begin{array}{rcl} \text{1st card black} & \longrightarrow & \text{1st plot gets } B \\ \text{2nd card red} & \longrightarrow & \text{2nd plot gets } A \\ \text{3rd card black} & \longrightarrow & \text{3rd plot gets } B \\ & \vdots & \end{array}$$

This is the first design we will study, a **completely randomized design**.

Results

B	*A*	*B*	*A*	*B*	*B*
26.9	11.4	26.6	23.7	25.3	28.5
B	*A*	*A*	*A*	*B*	*A*
14.2	17.9	16.5	21.1	24.3	19.6

How much evidence is there that fertilizer type is a source of yield variation?

Evidence about differences between two populations is generally measured by comparing summary statistics across the two sample populations.

(Recall, a **statistic** is any computable function of known, observed data).

Results

B	*A*	*B*	*A*	*B*	*B*
26.9	11.4	26.6	23.7	25.3	28.5
B	*A*	*A*	*A*	*B*	*A*
14.2	17.9	16.5	21.1	24.3	19.6

How much evidence is there that fertilizer type is a source of yield variation?

Evidence about differences between two populations is generally measured by comparing summary statistics across the two sample populations.

(Recall, a **statistic** is any computable function of known, observed data).

Results

B	*A*	*B*	*A*	*B*	*B*
26.9	11.4	26.6	23.7	25.3	28.5
B	*A*	*A*	*A*	*B*	*A*
14.2	17.9	16.5	21.1	24.3	19.6

How much evidence is there that fertilizer type is a source of yield variation?

Evidence about differences between two populations is generally measured by comparing summary statistics across the two sample populations.

(Recall, a **statistic** is any computable function of known, observed data).

Summaries of sample distribution

- Empirical distribution: $\hat{\Pr}(a, b] = \#(a < y_i \leq b)/n$
- Empirical CDF (cumulative distribution function)

$$\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$$

- Histograms
- Kernel density estimates

These summaries retain all the data information except the unit labels.

Summaries of sample distribution

- Empirical distribution: $\hat{\Pr}(a, b] = \#(a < y_i \leq b)/n$
- Empirical CDF (cumulative distribution function)

$$\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$$

- Histograms
- Kernel density estimates

These summaries retain all the data information except the unit labels.

Summaries of sample distribution

- Empirical distribution: $\hat{\Pr}(a, b] = \#(a < y_i \leq b)/n$
- Empirical CDF (cumulative distribution function)

$$\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$$

- Histograms
- Kernel density estimates

These summaries retain all the data information except the unit labels.

Summaries of sample distribution

- Empirical distribution: $\hat{\Pr}(a, b] = \#(a < y_i \leq b)/n$
- Empirical CDF (cumulative distribution function)

$$\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$$

- Histograms
- Kernel density estimates

These summaries retain all the data information except the unit labels.

Summaries of sample distribution

- Empirical distribution: $\hat{\Pr}(a, b] = \#(a < y_i \leq b)/n$
- Empirical CDF (cumulative distribution function)

$$\hat{F}(y) = \#(y_i \leq y)/n = \hat{\Pr}(-\infty, y]$$

- Histograms
- Kernel density estimates

These summaries retain all the data information except the unit labels.

Summaries of sample location

- sample mean or average : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- sample median : A/the value $y_{.5}$ such that

$$\frac{\#(y_i \leq y_{.5})}{n} \geq 1/2 \qquad \frac{\#(y_i \geq y_{.5})}{n} \geq 1/2$$

To find the median, sort the data in increasing order, and call these values $y_{(1)}, \ldots, y_{(n)}$. If there are no ties, then

if n is odd, then $y_{(\frac{n+1}{2})}$ is the median;

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}$$

if n is even, then all numbers between $y_{(\frac{n}{2})}$ and $y_{(\frac{n}{2}+1)}$ are medians.

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}, y_{(8)}$$

Summaries of sample location

- sample mean or average : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- sample median : A/the value $y_{.5}$ such that

$$\frac{\#(y_i \leq y_{.5})}{n} \geq 1/2 \qquad \frac{\#(y_i \geq y_{.5})}{n} \geq 1/2$$

To find the median, sort the data in increasing order, and call these values $y_{(1)}, \ldots, y_{(n)}$. If there are no ties, then

if n is odd, then $y_{(\frac{n+1}{2})}$ is the median;

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}$$

if n is even, then all numbers between $y_{(\frac{n}{2})}$ and $y_{(\frac{n}{2}+1)}$ are medians.

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}, y_{(8)}$$

Summaries of sample location

- sample mean or average : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- sample median : A/the value $y_{.5}$ such that

$$\frac{\#(y_i \leq y_{.5})}{n} \geq 1/2 \qquad \frac{\#(y_i \geq y_{.5})}{n} \geq 1/2$$

To find the median, sort the data in increasing order, and call these values $y_{(1)}, \ldots, y_{(n)}$. If there are no ties, then

if n is odd, then $y_{(\frac{n+1}{2})}$ is the median;

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}$$

if n is even, then all numbers between $y_{(\frac{n}{2})}$ and $y_{(\frac{n}{2}+1)}$ are medians.

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}, y_{(8)}$$

Summaries of sample location

- sample mean or average : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- sample median : A/the value $y_{.5}$ such that

$$\frac{\#(y_i \leq y_{.5})}{n} \geq 1/2 \qquad \frac{\#(y_i \geq y_{.5})}{n} \geq 1/2$$

To find the median, sort the data in increasing order, and call these values $y_{(1)}, \ldots, y_{(n)}$. If there are no ties, then

if n is odd, then $y_{(\frac{n+1}{2})}$ is the median;

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}$$

if n is even, then all numbers between $y_{(\frac{n}{2})}$ and $y_{(\frac{n}{2}+1)}$ are medians.

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}, y_{(8)}$$

Summaries of sample location

- sample mean or average : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- sample median : A/the value $y_{.5}$ such that

$$\frac{\#(y_i \leq y_{.5})}{n} \geq 1/2 \qquad \frac{\#(y_i \geq y_{.5})}{n} \geq 1/2$$

To find the median, sort the data in increasing order, and call these values $y_{(1)}, \ldots, y_{(n)}$. If there are no ties, then

if n is odd, then $y_{(\frac{n+1}{2})}$ is the median;

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}$$

if n is even, then all numbers between $y_{(\frac{n}{2})}$ and $y_{(\frac{n}{2}+1)}$ are medians.

$$y_{(1)}, y_{(2)}, y_{(3)}, y_{(4)}, y_{(5)}, y_{(6)}, y_{(7)}, y_{(8)}$$

Summaries of sample scale

- sample variance and standard deviation:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s = \sqrt{s^2}$$

- interquantile range:

  $[y_{.25}, y_{.75}]$ (interquartile range)
  $[y_{.025}, y_{.975}]$ (95% interval)

Summaries of sample scale

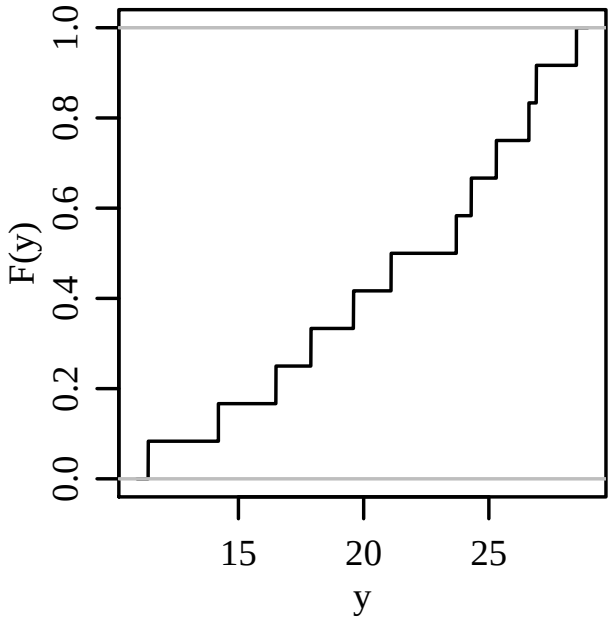
- sample variance and standard deviation:

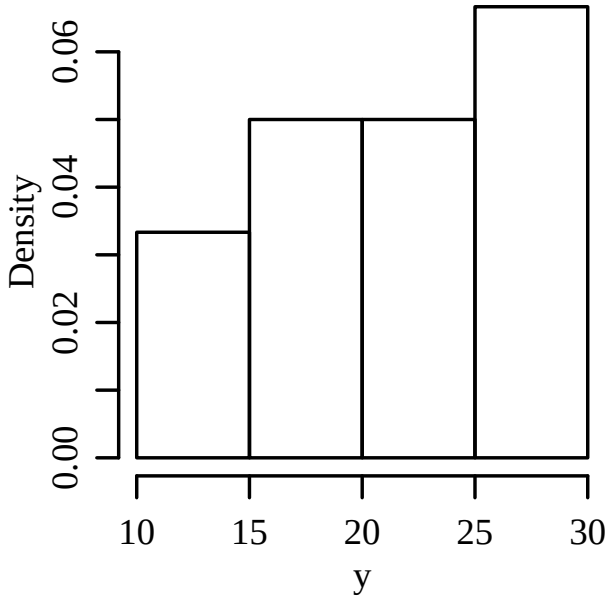
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s = \sqrt{s^2}$$

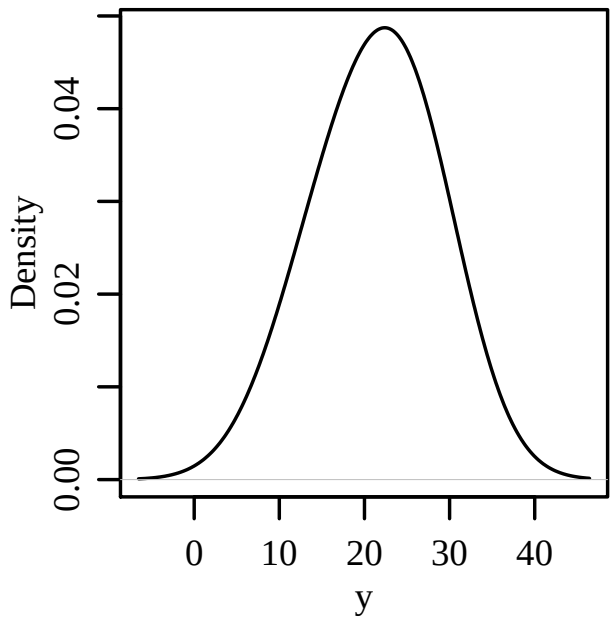
- interquantile range:

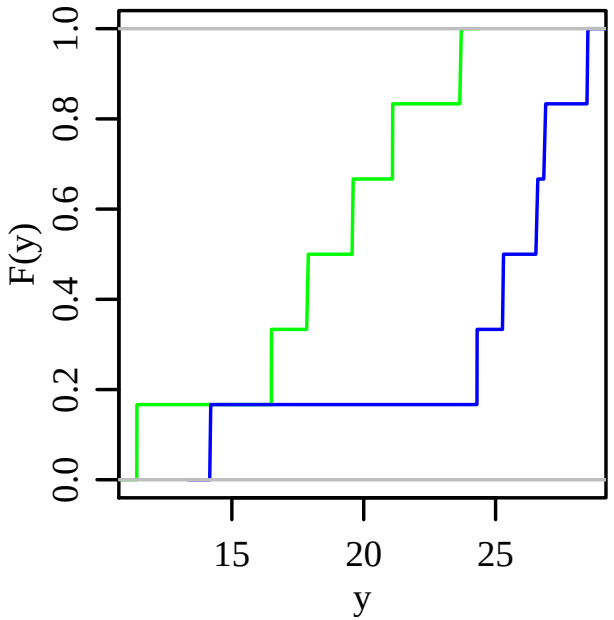
  $[y_{.25}, y_{.75}]$ (interquartile range)
  $[y_{.025}, y_{.975}]$ (95% interval)

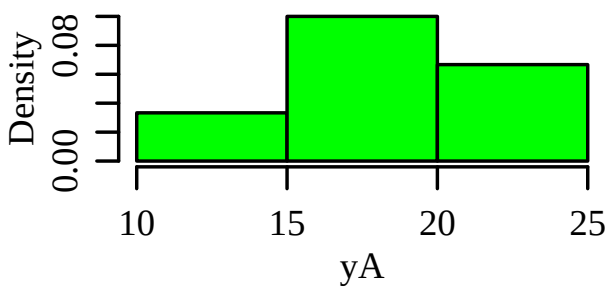
Example: Wheat yield

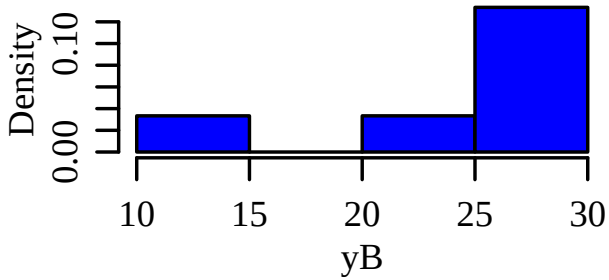


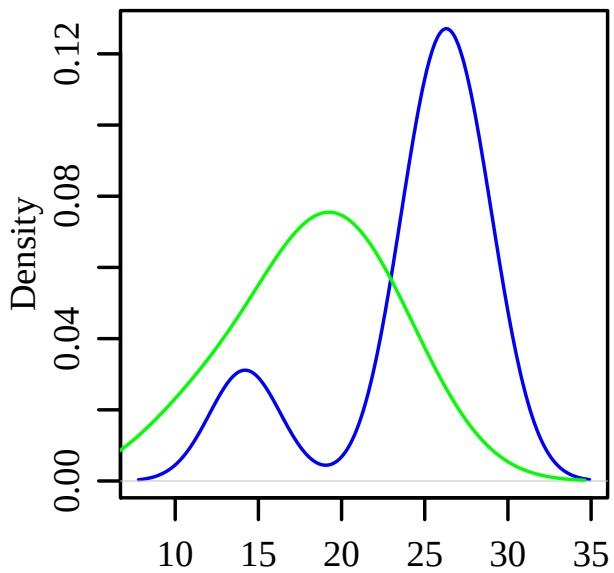












Summaries in R

All of these sample summaries are easily obtained in R:

```
> yA<-c(11.4, 23.7, 17.9, 16.5, 21.1, 19.6)
> yB<-c(26.9, 26.6, 25.3, 28.5, 14.2, 24.3)


> mean(yA)
[1] 18.36667
> mean(yB)
[1] 24.3

> median(yA)
[1] 18.75
> median(yB)
[1] 25.95

> sd(yA)
[1] 4.234934
> sd(yB)
[1] 5.151699

> quantile(yA,prob=c(.25,.75))
   25%    75%
16.850 20.725
> quantile(yB,prob=c(.25,.75))
   25%    75%
24.550 26.825
```

Induction and generalization

So there is a difference in yield for *these* wheat fields.

Would you recommend B over A for future plantings?

Do you think these results generalize to a *larger population*?

Induction and generalization

So there is a difference in yield for *these* wheat fields.

Would you recommend B over A for future plantings?

Do you think these results generalize to a *larger population*?

Induction and generalization

So there is a difference in yield for *these* wheat fields.

Would you recommend B over A for future plantings?

Do you think these results generalize to a *larger population*?

Hypotheses: competing explanations

Questions:

- Could the observed differences be due to fertilizer type?
- Could the observed differences be due to plot-to-plot variation?

Hypothesis tests:

- H_0 (null hypothesis): Fertilizer type does not affect yield.
- H_1 (alternative hypothesis): Fertilizer type does affect yield.

A **statistical hypothesis test** evaluates the compatibility of H_0 with the data.

Hypotheses: competing explanations

Questions:

- Could the observed differences be due to fertilizer type?
- Could the observed differences be due to plot-to-plot variation?

Hypothesis tests:

- H_0 (null hypothesis): Fertilizer type does not affect yield.
- H_1 (alternative hypothesis): Fertilizer type does affect yield.

A **statistical hypothesis test** evaluates the compatibility of H_0 with the data.

Test statistics and null distributions

Suppose we are interested in mean wheat yields. We can evaluate H_0 by answering the following questions:

- Is a mean difference of 5.93 plausible/probable if H_0 is true?
- Is a mean difference of 5.93 large compared to experimental noise?

To answer the above, we need to compare

$\{|\bar{y}_B - \bar{y}_A| = 5.93\}$, the *observed difference* in the experiment

to

values of $|\bar{y}_B - \bar{y}_A|$ that *could have been observed if* H_0 *were true*.

Hypothetical values of $|\bar{y}_B - \bar{y}_A|$ that could have been observed under H_0 are referred to as samples from the **null distribution**.

Test statistics and null distributions

Suppose we are interested in mean wheat yields. We can evaluate H_0 by answering the following questions:

- Is a mean difference of 5.93 plausible/probable if H_0 is true?
- Is a mean difference of 5.93 large compared to experimental noise?

To answer the above, we need to compare

$$\{|\bar{y}_B - \bar{y}_A| = 5.93\}, \text{ the } \textit{observed difference} \text{ in the experiment}$$

to

values of $|\bar{y}_B - \bar{y}_A|$ that *could have been observed if* H_0 *were true*.

Hypothetical values of $|\bar{y}_B - \bar{y}_A|$ that could have been observed under H_0 are referred to as samples from the **null distribution**.

Test statistics and null distributions

Suppose we are interested in mean wheat yields. We can evaluate H_0 by answering the following questions:

- Is a mean difference of 5.93 plausible/probable if H_0 is true?
- Is a mean difference of 5.93 large compared to experimental noise?

To answer the above, we need to compare

$\{|\bar{y}_B - \bar{y}_A| = 5.93\}$, the *observed difference* in the experiment

to

values of $|\bar{y}_B - \bar{y}_A|$ that *could have been observed if* H_0 *were true*.

Hypothetical values of $|\bar{y}_B - \bar{y}_A|$ that could have been observed under H_0 are referred to as samples from the **null distribution**.

Test statistics and null distributions

Suppose we are interested in mean wheat yields. We can evaluate H_0 by answering the following questions:

- Is a mean difference of 5.93 plausible/probable if H_0 is true?
- Is a mean difference of 5.93 large compared to experimental noise?

To answer the above, we need to compare

$\{|\bar{y}_B - \bar{y}_A| = 5.93\}$, the *observed difference* in the experiment

to

values of $|\bar{y}_B - \bar{y}_A|$ that *could have been observed if* H_0 *were true*.

Hypothetical values of $|\bar{y}_B - \bar{y}_A|$ that could have been observed under H_0 are referred to as samples from the **null distribution**.

Test statistics and null distributions

$$g(\mathbf{Y}_A, \mathbf{Y}_B) = g(\{Y_{1,A}, \ldots, Y_{6,A}\}, \{Y_{1,B}, \ldots, Y_{6,B}\}) = |\bar{Y}_B - \bar{Y}_A|.$$

This is a function of the outcome of the experiment. It is a **statistic**.

Since we will use it to perform a hypothesis test, we will call it a **test statistic**.

Observed test statistic:

$g(11.4, 23.7, \ldots, 14.2, 24.3) = 5.93 = g_{\text{obs}}$

Hypothesis testing procedure:

Compare g_{obs} to $g(\mathbf{Y_A}, \mathbf{Y_B})$, where

$\mathbf{Y_A}$ and $\mathbf{Y_B}$ are values that *could have* been observed, if H_0 were true.

Test statistics and null distributions

$$g(\mathbf{Y}_A, \mathbf{Y}_B) = g(\{Y_{1,A}, \ldots, Y_{6,A}\}, \{Y_{1,B}, \ldots, Y_{6,B}\}) = |\bar{Y}_B - \bar{Y}_A|.$$

This is a function of the outcome of the experiment. It is a **statistic**.

Since we will use it to perform a hypothesis test, we will call it a **test statistic**.

Observed test statistic:

$g(11.4, 23.7, \ldots, 14.2, 24.3) = 5.93 = g_{\text{obs}}$

Hypothesis testing procedure:

Compare g_{obs} to $g(\mathbf{Y_A}, \mathbf{Y_B})$, where

$\mathbf{Y_A}$ and $\mathbf{Y_B}$ are values that *could have* been observed, if H_0 were true.

Test statistics and null distributions

$$g(\mathbf{Y}_A, \mathbf{Y}_B) = g(\{Y_{1,A}, \ldots, Y_{6,A}\}, \{Y_{1,B}, \ldots, Y_{6,B}\}) = |\bar{Y}_B - \bar{Y}_A|.$$

This is a function of the outcome of the experiment. It is a **statistic**.

Since we will use it to perform a hypothesis test, we will call it a **test statistic**.

Observed test statistic:

$g(11.4, 23.7, \ldots, 14.2, 24.3) = 5.93 = g_{\text{obs}}$

Hypothesis testing procedure:

Compare g_{obs} to $g(\mathbf{Y_A}, \mathbf{Y_B})$, where

$\mathbf{Y_A}$ and $\mathbf{Y_B}$ are values that *could have* been observed, if H_0 were true.

Experimental procedure and observed outcome

Recall the design of the experiment:

1. Shuffled cards were dealt $B, R, B, R, \ldots$, fertilizers assigned to subplots:

B	A	B	A	B	B
B	A	A	A	B	A

2. Crops were grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	A 23.7	B 25.3	B 28.5
B 14.2	A 17.9	A 16.5	A 21.1	B 24.3	A 19.6

Experimental procedure and observed outcome

Recall the design of the experiment:

1. Shuffled cards were dealt $B, R, B, R, \ldots$, fertilizers assigned to subplots:

B	A	B	A	B	B
B	A	A	A	B	A

2. Crops were grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	A 23.7	B 25.3	B 28.5
B 14.2	A 17.9	A 16.5	A 21.1	B 24.3	A 19.6

Experimental procedure and observed outcome

Recall the design of the experiment:

1. Shuffled cards were dealt $B, R, B, R, \ldots$, fertilizers assigned to subplots:

B	A	B	A	B	B
B	A	A	A	B	A

2. Crops were grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	A 23.7	B 25.3	B 28.5
B 14.2	A 17.9	A 16.5	A 21.1	B 24.3	A 19.6

Experimental procedure and observed outcome

Recall the design of the experiment:

1. Shuffled cards were dealt $B, R, B, R, \ldots$, fertilizers assigned to subplots:

B	A	B	A	B	B
B	A	A	A	B	A

2. Crops were grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	A 23.7	B 25.3	B 28.5
B 14.2	A 17.9	A 16.5	A 21.1	B 24.3	A 19.6

Experimental procedure and potential outcome

Imagine **re-doing** the experiment if "H_0: no treatment effect" were true:

1. Shuffled cards were dealt $B, R, B, B, \ldots$, fertilizers assigned to subplots:

B	A	B	B	A	A
A	B	B	A	A	B

2. Crops are grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	B 23.7	A 25.3	A 28.5
A 14.2	B 17.9	B 16.5	A 21.1	A 24.3	B 19.6

Under this hypothetical treatment assignment,

$$\begin{aligned} (\mathbf{Y}_A, \mathbf{Y}_B) &= \{11.4, 25.3, \ldots, 21.1, 19.6\} \\ |\bar{Y}_B - \bar{Y}_A| &= 1.07 \end{aligned}$$

This represents an outcome of the experiment in a universe where

- The treatment assignment is $B, A, B, B, A, A, A, B, B, A, A, B$;
- H_0 is true.

Experimental procedure and potential outcome

Imagine **re-doing** the experiment if "H_0: no treatment effect" were true:

1. Shuffled cards were dealt $B, R, B, B, \ldots$, fertilizers assigned to subplots:

B	A	B	B	A	A
A	B	B	A	A	B

2. Crops are grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	B 23.7	A 25.3	A 28.5
A 14.2	B 17.9	B 16.5	A 21.1	A 24.3	B 19.6

Under this hypothetical treatment assignment,

$$\begin{aligned} (\mathbf{Y}_A, \mathbf{Y}_B) &= \{11.4, 25.3, \ldots, 21.1, 19.6\} \\ |\bar{Y}_B - \bar{Y}_A| &= 1.07 \end{aligned}$$

This represents an outcome of the experiment in a universe where

- The treatment assignment is $B, A, B, B, A, A, A, B, B, A, A, B$;
- H_0 is true.

Experimental procedure and potential outcome

Imagine **re-doing** the experiment if "H_0: no treatment effect" were true:

1. Shuffled cards were dealt $B, R, B, B, \ldots$, fertilizers assigned to subplots:

B	A	B	B	A	A
A	B	B	A	A	B

2. Crops are grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	B 23.7	A 25.3	A 28.5
A 14.2	B 17.9	B 16.5	A 21.1	A 24.3	B 19.6

Under this hypothetical treatment assignment,

$$\begin{aligned} (\mathbf{Y}_A, \mathbf{Y}_B) &= \{11.4, 25.3, \ldots, 21.1, 19.6\} \\ |\bar{Y}_B - \bar{Y}_A| &= 1.07 \end{aligned}$$

This represents an outcome of the experiment in a universe where

- The treatment assignment is $B, A, B, B, A, A, A, B, B, A, A, B$;
- H_0 is true.

Experimental procedure and potential outcome

Imagine **re-doing** the experiment if "H_0: no treatment effect" were true:

1. Shuffled cards were dealt $B, R, B, B, \ldots$, fertilizers assigned to subplots:

B	A	B	B	A	A
A	B	B	A	A	B

2. Crops are grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	B 23.7	A 25.3	A 28.5
A 14.2	B 17.9	B 16.5	A 21.1	A 24.3	B 19.6

Under this hypothetical treatment assignment,

$$\begin{aligned} (\mathbf{Y}_A, \mathbf{Y}_B) &= \{11.4, 25.3, \ldots, 21.1, 19.6\} \\ |\bar{Y}_B - \bar{Y}_A| &= 1.07 \end{aligned}$$

This represents an outcome of the experiment in a universe where

- The treatment assignment is $B, A, B, B, A, A, A, B, B, A, A, B$;
- H_0 is true.

Experimental procedure and potential outcome

Imagine **re-doing** the experiment if "H_0: no treatment effect" were true:

1. Shuffled cards were dealt $B, R, B, B, \ldots$, fertilizers assigned to subplots:

B	A	B	B	A	A
A	B	B	A	A	B

2. Crops are grown and wheat yields obtained:

B 26.9	A 11.4	B 26.6	B 23.7	A 25.3	A 28.5
A 14.2	B 17.9	B 16.5	A 21.1	A 24.3	B 19.6

Under this hypothetical treatment assignment,

$$\begin{aligned} (\mathbf{Y}_A, \mathbf{Y}_B) &= \{11.4, 25.3, \ldots, 21.1, 19.6\} \\ |\bar{Y}_B - \bar{Y}_A| &= 1.07 \end{aligned}$$

This represents an outcome of the experiment in a universe where

- The treatment assignment is $B, A, B, B, A, A, A, B, B, A, A, B$;
- H_0 is true.

The null distribution

IDEA:

To consider what types of outcomes we would see in universes where H_0 is true, compute $g(\mathbf{Y}_A, \mathbf{Y}_B)$ for each possible treatment assignment, assuming H_0 true.

Under our randomization scheme, there were

$$\frac{12!}{6!6!} = \binom{12}{6} = 924$$

equally likely ways the treatments could have been assigned.

For each one of these, we can calculate the value of the test statistic that would've been observed under H_0:

$$\{g_1, g_2, \ldots, g_{924}\}$$

The null distribution

IDEA:

To consider what types of outcomes we would see in universes where H_0 is true, compute $g(\mathbf{Y}_A, \mathbf{Y}_B)$ for each possible treatment assignment, assuming H_0 true.

Under our randomization scheme, there were

$$\frac{12!}{6!6!} = \binom{12}{6} = 924$$

equally likely ways the treatments could have been assigned.

For each one of these, we can calculate the value of the test statistic that would've been observed under H_0:

$$\{g_1, g_2, \ldots, g_{924}\}$$

The null distribution

IDEA:

To consider what types of outcomes we would see in universes where H_0 is true, compute $g(\mathbf{Y}_A, \mathbf{Y}_B)$ for each possible treatment assignment, assuming H_0 true.

Under our randomization scheme, there were

$$\frac{12!}{6!6!} = \binom{12}{6} = 924$$

equally likely ways the treatments could have been assigned.

For each one of these, we can calculate the value of the test statistic that would've been observed under H_0:

$$\{g_1, g_2, \ldots, g_{924}\}$$

The null distribution

$$\{g_1, g_2, \ldots, g_{924}\}$$

This enumerates *all potential* pre-randomization outcomes of our test statistic, assuming *no treatment effect*.

Since each treatment assignment was equally likely, these values give a **null distribution**: a probability distribution of possible experimental results, if H_0 were true.

$$F(x|\mathrm{H}_0) = \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \leq x|\mathrm{H}_0) = \frac{\#\{g_k \leq x\}}{924}$$

This distribution is sometimes called the **randomization distribution**, because it is obtained by the randomization scheme of the experiment.

The null distribution

$$\{g_1, g_2, \ldots, g_{924}\}$$

This enumerates *all potential* pre-randomization outcomes of our test statistic, assuming *no treatment effect*.

Since each treatment assignment was equally likely, these values give a **null distribution**: a probability distribution of possible experimental results, if H_0 were true.

$$F(x|\mathrm{H}_0) = \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \leq x|\mathrm{H}_0) = \frac{\#\{g_k \leq x\}}{924}$$

This distribution is sometimes called the **randomization distribution**, because it is obtained by the randomization scheme of the experiment.

The null distribution

$$\{g_1, g_2, \ldots, g_{924}\}$$

This enumerates *all potential* pre-randomization outcomes of our test statistic, assuming *no treatment effect*.

Since each treatment assignment was equally likely, these values give a **null distribution**: a probability distribution of possible experimental results, if H_0 were true.

$$F(x|\mathrm{H}_0) = \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \leq x|\mathrm{H}_0) = \frac{\#\{g_k \leq x\}}{924}$$

This distribution is sometimes called the **randomization distribution**, because it is obtained by the randomization scheme of the experiment.

Null distribution, wheat example

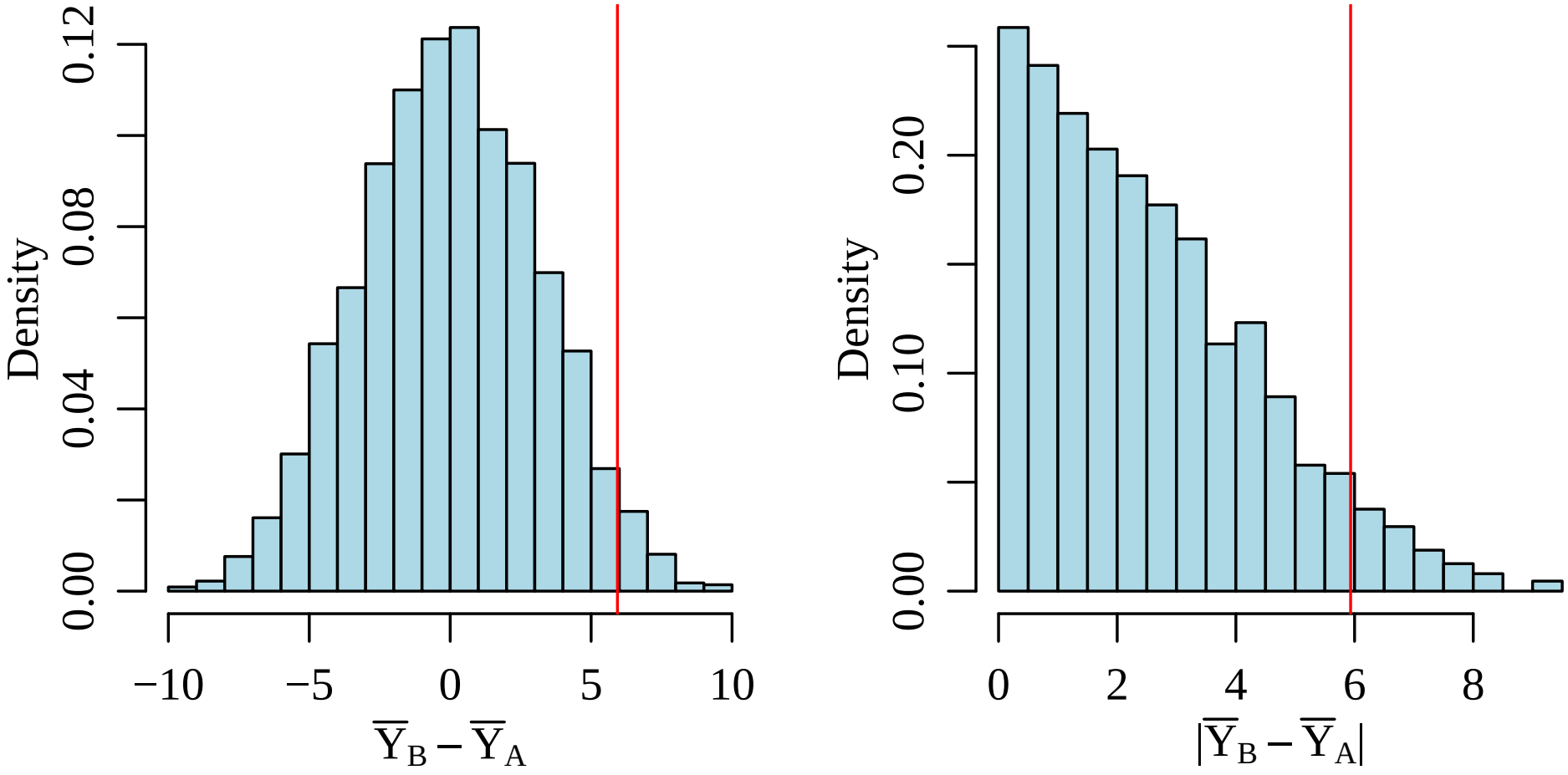


Figure: Approximate randomization distribution for the wheat example

Comparing data to the null distribution:

Is there any contradiction between H_0 and our data?

$$\Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq 5.93 | \mathrm{H}_0) = 0.056$$

The probability of observing a difference of 5.93 or more is unlikely under H_0.

This probability calculation is called a **p-value**. Generically, a *p*-value is

> "The probability, under the null hypothesis, of obtaining a result as or more extreme than the observed result."

The basic idea:

small *p*-value $\rightarrow$ evidence against H_0
large *p*-value $\rightarrow$ no evidence against H_0

Comparing data to the null distribution:

Is there any contradiction between H_0 and our data?

$$\Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq 5.93 | \mathrm{H}_0) = 0.056$$

The probability of observing a difference of 5.93 or more is unlikely under H_0.

This probability calculation is called a **p-value**. Generically, a *p*-value is

> "The probability, under the null hypothesis, of obtaining a result as or more extreme than the observed result."

The basic idea:

small *p*-value $\rightarrow$ evidence against H_0
large *p*-value $\rightarrow$ no evidence against H_0

Comparing data to the null distribution:

Is there any contradiction between H_0 and our data?

$$\Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq 5.93 | \mathrm{H}_0) = 0.056$$

The probability of observing a difference of 5.93 or more is unlikely under H_0.

This probability calculation is called a **p-value**. Generically, a *p*-value is

"The probability, under the null hypothesis, of obtaining a result as or more extreme than the observed result."

The basic idea:

small *p*-value $\rightarrow$ evidence against H_0
large *p*-value $\rightarrow$ no evidence against H_0

Comparing data to the null distribution:

Is there any contradiction between H_0 and our data?

$$\Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq 5.93 | \mathrm{H}_0) = 0.056$$

The probability of observing a difference of 5.93 or more is unlikely under H_0.

This probability calculation is called a **p-value**. Generically, a *p*-value is

"The probability, under the null hypothesis, of obtaining a result as or more extreme than the observed result."

The basic idea:

small *p*-value	$\longrightarrow$	evidence against H_0
large *p*-value	$\longrightarrow$	no evidence against H_0

Approximating a randomization distribution:

We don't want to have to enumerate all $\binom{n_A+n_B}{n_A}$ possible treatment assignments. Instead, repeat the following S times for some large number S:

(a) randomly simulate a treatment assignment from the population of possible treatment assignments, under the randomization scheme.

(b) compute the value of the test statistic, given the simulated treatment assignment and under H_0.

The **empirical distribution** of $\{g_1, \ldots, g_S\}$ **approximates** the null distribution:

$$\frac{\#(g_s \geq g_{\text{obs}})}{S} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq g_{\text{obs}} | \mathrm{H}_0)$$

The approximation improves if S is increased.

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c("B", "A", "B", "A", "B", "B", "B", "A", "A", "A", "B", "A")

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim=="B"]) - mean(y[xsim=="A"] ) )
}
```

Approximating a randomization distribution:

We don't want to have to enumerate all $\binom{n_A+n_B}{n_A}$ possible treatment assignments. Instead, repeat the following S times for some large number S:

(a) randomly simulate a treatment assignment from the population of possible treatment assignments, under the randomization scheme.

(b) compute the value of the test statistic, given the simulated treatment assignment and under H_0.

The **empirical distribution** of $\{g_1, \ldots, g_S\}$ **approximates** the null distribution:

$$\frac{\#(g_s \geq g_{\text{obs}})}{S} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq g_{\text{obs}} | \mathrm{H}_0)$$

The approximation improves if S is increased.

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c("B", "A", "B", "A", "B", "B", "B", "A", "A", "A", "B", "A")

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim=="B"]) - mean(y[xsim=="A"] ) )
}
```

Approximating a randomization distribution:

We don't want to have to enumerate all $\binom{n_A+n_B}{n_A}$ possible treatment assignments. Instead, repeat the following S times for some large number S:

(a) randomly simulate a treatment assignment from the population of possible treatment assignments, under the randomization scheme.

(b) compute the value of the test statistic, given the simulated treatment assignment and under H_0.

The **empirical distribution** of $\{g_1, \ldots, g_S\}$ **approximates** the null distribution:

$$\frac{\#(g_s \geq g_{\text{obs}})}{S} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq g_{\text{obs}} | \mathrm{H}_0)$$

The approximation improves if S is increased.

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c("B", "A", "B", "A", "B", "B", "B", "A", "A", "A", "B", "A")

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim=="B"]) - mean(y[xsim=="A"] ) )
}
```

Approximating a randomization distribution:

We don't want to have to enumerate all $\binom{n_A+n_B}{n_A}$ possible treatment assignments. Instead, repeat the following S times for some large number S:

(a) randomly simulate a treatment assignment from the population of possible treatment assignments, under the randomization scheme.

(b) compute the value of the test statistic, given the simulated treatment assignment and under H_0.

The **empirical distribution** of $\{g_1, \ldots, g_S\}$ **approximates** the null distribution:

$$\frac{\#(g_s \geq g_{\text{obs}})}{S} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq g_{\text{obs}} | \mathrm{H}_0)$$

The approximation improves if S is increased.

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c("B", "A", "B", "A", "B", "B", "B", "A", "A", "A", "B", "A")

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim=="B"]) - mean(y[xsim=="A"] ) )
}
```

Approximating a randomization distribution:

We don't want to have to enumerate all $\binom{n_A+n_B}{n_A}$ possible treatment assignments. Instead, repeat the following S times for some large number S:

(a) randomly simulate a treatment assignment from the population of possible treatment assignments, under the randomization scheme.

(b) compute the value of the test statistic, given the simulated treatment assignment and under H_0.

The **empirical distribution** of $\{g_1, \ldots, g_S\}$ **approximates** the null distribution:

$$\frac{\#(g_s \geq g_{\text{obs}})}{S} \approx \Pr(g(\mathbf{Y}_A, \mathbf{Y}_B) \geq g_{\text{obs}} | \mathrm{H}_0)$$

The approximation improves if S is increased.

Here is some R-code:

```
y<-c(26.9,11.4,26.6,23.7,25.3,28.5,14.2,17.9,16.5,21.1,24.3,19.6)
x<-c("B", "A", "B", "A", "B", "B", "B", "A", "A", "A", "B", "A")

g.null<-real()
for(s in 1:10000)
{
  xsim<-sample(x)
  g.null[s]<- abs( mean(y[xsim=="B"]) - mean(y[xsim=="A"] ) )
}
```

Essential nature of a hypothesis test

Given H_0, H_1 and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant **test statistic** $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a **null distribution** : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under H_0. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that *could have happened under* H_0.
3. Compute the **p-value**: The probability under H_0 of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y}) | \mathrm{H}_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0
If the p-value is large $\Rightarrow$ not evidence against H_0

Even if we follow these guidelines, we must be careful in our specification of H_0, H_1 and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

Essential nature of a hypothesis test

Given H_0, H_1 and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant **test statistic** $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a **null distribution** : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under H_0. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that *could have happened under* H_0.

3. Compute the **p-value**: The probability under H_0 of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y}) | \mathrm{H}_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0
If the p-value is large $\Rightarrow$ not evidence against H_0

Even if we follow these guidelines, we must be careful in our specification of H_0, H_1 and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

Essential nature of a hypothesis test

Given H_0, H_1 and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant **test statistic** $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a **null distribution** : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under H_0. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that *could have happened under* H_0.
3. Compute the **p-value**: The probability under H_0 of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y}) | \mathrm{H}_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0
If the p-value is large $\Rightarrow$ not evidence against H_0

Even if we follow these guidelines, we must be careful in our specification of H_0, H_1 and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

Essential nature of a hypothesis test

Given H_0, H_1 and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant **test statistic** $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a **null distribution** : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under H_0. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that *could have happened under* H_0.
3. Compute the **p-value**: The probability under H_0 of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y}) | \mathrm{H}_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0

If the p-value is large $\Rightarrow$ not evidence against H_0

Even if we follow these guidelines, we must be careful in our specification of H_0, H_1 and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

Essential nature of a hypothesis test

Given H_0, H_1 and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant **test statistic** $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a **null distribution** : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under H_0. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that *could have happened under* H_0.
3. Compute the **p-value**: The probability under H_0 of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y}) | \mathrm{H}_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0
If the p-value is large $\Rightarrow$ not evidence against H_0

Even if we follow these guidelines, we must be careful in our specification of H_0, H_1 and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

Essential nature of a hypothesis test

Given H_0, H_1 and data $\mathbf{y} = \{y_1, \ldots, y_n\}$:

1. From the data, compute a relevant **test statistic** $g(\mathbf{y})$: The test statistic $g(\mathbf{y})$ should be chosen so that it can differentiate between H_0 and H_1 in ways that are scientifically relevant. Typically, $g(\mathbf{y})$ is chosen so that

$$g(\mathbf{y}) \text{ is probably } \begin{cases} \text{small under } \mathrm{H}_0 \\ \text{large under } \mathrm{H}_1 \end{cases}$$

2. Obtain a **null distribution** : A probability distribution over the possible outcomes of $g(\mathbf{Y})$ under H_0. Here, $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ are potential experimental results that *could have happened under* H_0.
3. Compute the **p-value**: The probability under H_0 of observing a test statistic $g(\mathbf{Y})$ as or more extreme than the observed statistic $g(\mathbf{y})$.

$$p\text{-value} = \Pr(g(\mathbf{Y}) \geq g(\mathbf{y}) | \mathrm{H}_0)$$

If the p-value is small $\Rightarrow$ evidence against H_0
If the p-value is large $\Rightarrow$ not evidence against H_0

Even if we follow these guidelines, we must be careful in our specification of H_0, H_1 and $g(\mathbf{Y})$ for the hypothesis testing procedure to be useful.

Questions

- Is a small *p*-value evidence in favor of H_1?
- Is a large *p*-value evidence in favor of H_0?
- What does the *p*-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out.

Questions

- Is a small *p*-value evidence in favor of H_1?
- Is a large *p*-value evidence in favor of H_0?
- What does the *p*-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out.

Questions

- Is a small *p*-value evidence in favor of H_1?
- Is a large *p*-value evidence in favor of H_0?
- What does the *p*-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out.

Choosing test statistics

The test statistic $g(\mathbf{y})$ should be able to

"differentiate" between H_0 and H_1

in ways that are "scientifically relevant".

What does this mean?

Suppose our data consist of samples $\mathbf{y}_A$ and $\mathbf{y}_B$ from two populations A and B.

Previously we used $g(\mathbf{y}_A, \mathbf{y}_B) = |\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$.

Let's consider two different test statistics:

- t-statistic:
- Kolmogorov-Smirnov statistic

Choosing test statistics

The test statistic $g(\mathbf{y})$ should be able to

"differentiate" between H_0 and H_1

in ways that are "scientifically relevant".

What does this mean?

Suppose our data consist of samples $\mathbf{y}_A$ and $\mathbf{y}_B$ from two populations A and B.

Previously we used $g(\mathbf{y}_A, \mathbf{y}_B) = |\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$.

Let's consider two different test statistics:

- t-statistic:
- Kolmogorov-Smirnov statistic

Choosing test statistics

The test statistic $g(\mathbf{y})$ should be able to

"differentiate" between H_0 and H_1

in ways that are "scientifically relevant".

What does this mean?

Suppose our data consist of samples $\mathbf{y}_A$ and $\mathbf{y}_B$ from two populations A and B.

Previously we used $g(\mathbf{y}_A, \mathbf{y}_B) = |\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$.

Let's consider two different test statistics:

- t-statistic:
- Kolmogorov-Smirnov statistic

Choosing test statistics

The test statistic $g(\mathbf{y})$ should be able to

"differentiate" between H_0 and H_1

in ways that are "scientifically relevant".

What does this mean?

Suppose our data consist of samples $\mathbf{y}_A$ and $\mathbf{y}_B$ from two populations A and B.

Previously we used $g(\mathbf{y}_A, \mathbf{y}_B) = |\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$.

Let's consider two different test statistics:

- t-statistic:
- Kolmogorov-Smirnov statistic

The t statistic

$$g_t(\mathbf{y}_A, \mathbf{y}_B) = \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where}$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The t statistic

$$
\begin{aligned}
g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\
s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2
\end{aligned}
$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The t statistic

$$
\begin{aligned}
g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\
s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2
\end{aligned}
$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The t statistic

$$
\begin{aligned}
g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\
s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2
\end{aligned}
$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The t statistic

$$
\begin{aligned}
g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\
s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2
\end{aligned}
$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The t statistic

$$
\begin{aligned}
g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\
s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2
\end{aligned}
$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The t statistic

$$
\begin{aligned}
g_t(\mathbf{y}_A, \mathbf{y}_B) &= \frac{|\bar{y}_B - \bar{y}_A|}{s_p\sqrt{1/n_A + 1/n_B}}, \quad \text{where} \\
s_p^2 &= \frac{n_A - 1}{(n_A - 1) + (n_B - 1)} s_A^2 + \frac{n_B - 1}{(n_A - 1) + (n_B - 1)} s_B^2
\end{aligned}
$$

This is a scaled version of our previous test statistic, comparing the difference in sample means to a pooled version of the sample standard deviation and the sample size.

numerator: The effect size estimate (difference in means)

denominator: The precision of the estimate (sample sd scaled by sample size)

This statistic is

- increasing in $|\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_A|$;
- increasing in n_A and n_B;
- decreasing in s_p.

A more complete motivation for this statistic will be given in the next chapter.

The Kolmogorov-Smirnov statistic

$$g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = \max_{y \in \mathbb{R}} |\hat{F}_B(y) - \hat{F}_A(y)|$$

This is just the size of the largest gap between the two sample CDFs.

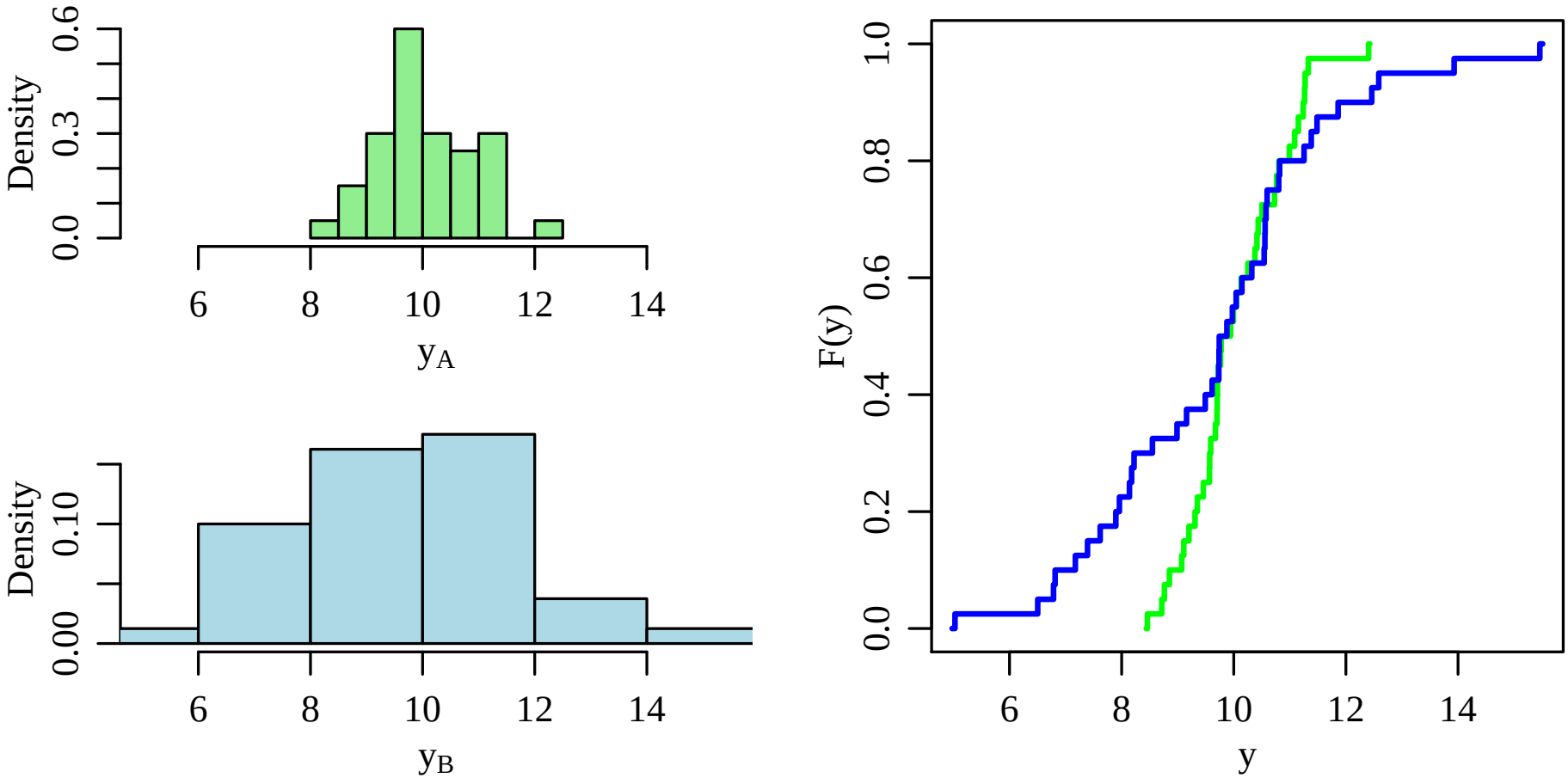


Figure: Histograms and empirical CDFs of the first two hypothetical samples.

Comparing the test statistics

Suppose we perform the CRD and obtain samples $\mathbf{y}_A$ and $\mathbf{y}_B$ given in Figure 3.

- $n_A = n_B = 40$
- $\bar{y}_A = 10.05$, $\bar{y}_B = 9.70$.
- $s_A = 0.87$, $s_B = 2.07$

The main difference seems to be the variances and not the means.

Hypothesis testing H_0: treatment does not affect response

We can approximate the null distributions of $g_t(\mathbf{Y}_A, \mathbf{Y}_B)$ and $g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B)$ by randomly reassigning the treatments but leaving the responses fixed:

```
Gsim<-NULL
for(s in 1:5000)
{
  xsim<-sample(x)
  yAsim<-y[xsim=="A"] ; yBsim<-y[xsim=="B"]
  g1<- g.tstat(yAsim,yBsim)
  g2<- g.ks(yAsim,yBsim)
  Gsim<-rbind(Gsim,c(g1,g2))
}
```

Comparing the test statistics

Suppose we perform the CRD and obtain samples $\mathbf{y}_A$ and $\mathbf{y}_B$ given in Figure 3.

- $n_A = n_B = 40$
- $\bar{y}_A = 10.05$, $\bar{y}_B = 9.70$.
- $s_A = 0.87$, $s_B = 2.07$

The main difference seems to be the variances and not the means.

Hypothesis testing H_0: treatment does not affect response

We can approximate the null distributions of $g_t(\mathbf{Y}_A, \mathbf{Y}_B)$ and $g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B)$ by randomly reassigning the treatments but leaving the responses fixed:

```
Gsim<-NULL
for(s in 1:5000)
{
  xsim<-sample(x)
  yAsim<-y[xsim=="A"] ; yBsim<-y[xsim=="B"]
  g1<- g.tstat(yAsim,yBsim)
  g2<- g.ks(yAsim,yBsim)
  Gsim<-rbind(Gsim,c(g1,g2))
}
```

Comparing the test statistics

Suppose we perform the CRD and obtain samples $\mathbf{y}_A$ and $\mathbf{y}_B$ given in Figure 3.

- $n_A = n_B = 40$
- $\bar{y}_A = 10.05$, $\bar{y}_B = 9.70$.
- $s_A = 0.87$, $s_B = 2.07$

The main difference seems to be the variances and not the means.

Hypothesis testing H_0: treatment does not affect response

We can approximate the null distributions of $g_t(\mathbf{Y}_A, \mathbf{Y}_B)$ and $g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B)$ by randomly reassigning the treatments but leaving the responses fixed:

```
Gsim<-NULL
for(s in 1:5000)
{
  xsim<-sample(x)
  yAsim<-y[xsim=="A"] ; yBsim<-y[xsim=="B"]
  g1<- g.tstat(yAsim,yBsim)
  g2<- g.ks(yAsim,yBsim)
  Gsim<-rbind(Gsim,c(g1,g2))
}
```

Comparing the test statistics

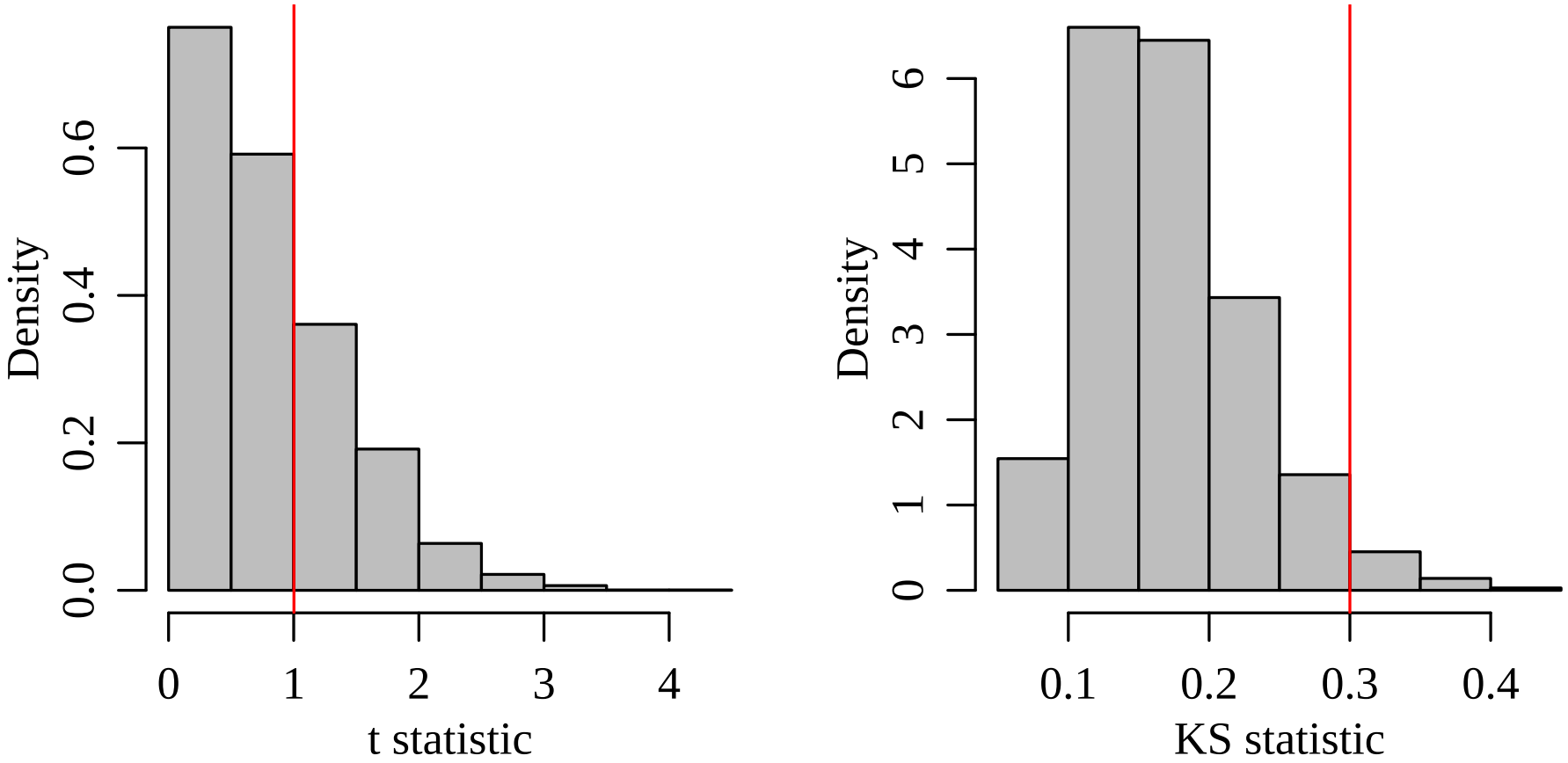


Figure: Randomization distributions for the t and KS statistics for the first example.

Comparing the test statistics

***p*-values**:

t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.00$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.00) = 0.321$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.30$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.30) = 0.043$

- test based on the *t*-statistic does not indicate strong evidence against H_0
- test based on the KS-statistic does.

Reason:

- The *t*-statistic is only sensitive to differences in means.
  In particular, if $\bar{y}_A = \bar{y}_B$ then the *t*-statistic is zero, its minimum value.
- In contrast, the KS-statistic is
  sensitive to any differences in the sample distributions.

Comparing the test statistics

***p*-values**:

t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.00$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.00) = 0.321$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.30$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.30) = 0.043$

- test based on the t-statistic does not indicate strong evidence against H_0
- test based on the KS-statistic does.

Reason:

- The t-statistic is only sensitive to differences in means.
  In particular, if $\bar{y}_A = \bar{y}_B$ then the t-statistic is zero, its minimum value.
- In contrast, the KS-statistic is
  sensitive to any differences in the sample distributions.

Comparing the test statistics

***p*-values**:

t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.00$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.00) = 0.321$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.30$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.30) = 0.043$

- test based on the t-statistic does not indicate strong evidence against H_0
- test based on the KS-statistic does.

Reason:

- The t-statistic is only sensitive to differences in means.
  In particular, if $\bar{y}_A = \bar{y}_B$ then the t-statistic is zero, its minimum value.
- In contrast, the KS-statistic is
  sensitive to any differences in the sample distributions.

Comparing the test statistics

***p*-values**:

t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.00$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.00) = 0.321$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.30$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.30) = 0.043$

- test based on the t-statistic does not indicate strong evidence against H_0
- test based on the KS-statistic does.

Reason:

- The t-statistic is only sensitive to differences in means.
  In particular, if $\bar{y}_A = \bar{y}_B$ then the t-statistic is zero, its minimum value.
- In contrast, the KS-statistic is
  sensitive to any differences in the sample distributions.

Comparing the test statistics

***p*-values**:

t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.00$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.00) = 0.321$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.30$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.30) = 0.043$

- test based on the t-statistic does not indicate strong evidence against H_0
- test based on the KS-statistic does.

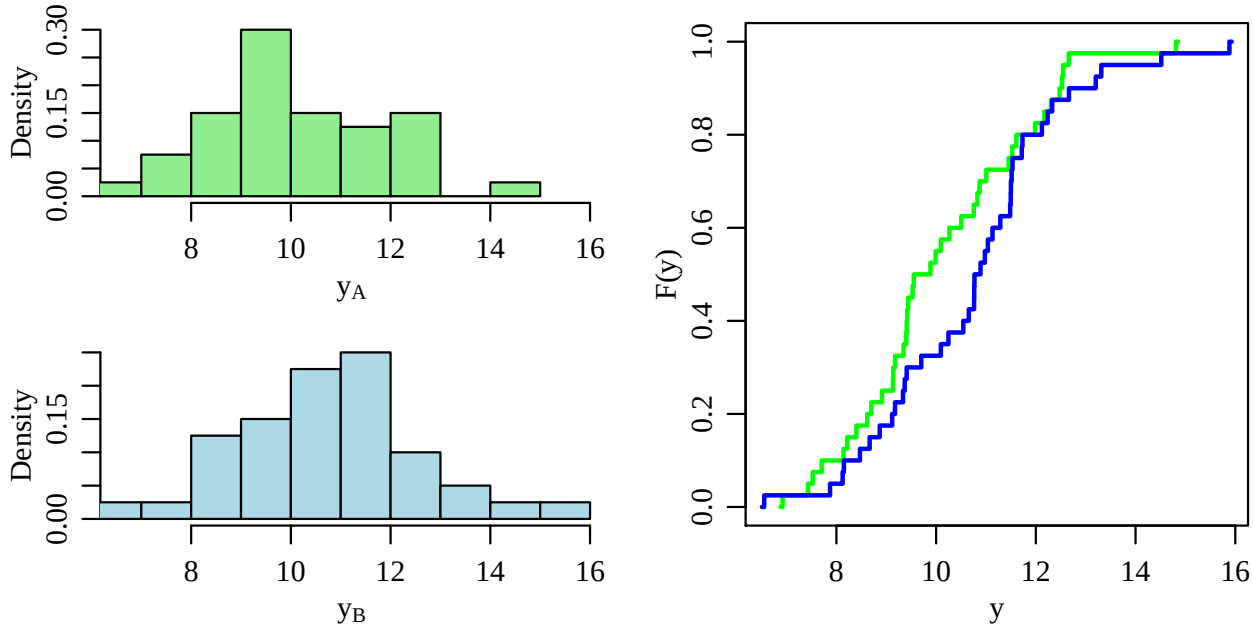
Reason:

- The t-statistic is only sensitive to differences in means.
  In particular, if $\bar{y}_A = \bar{y}_B$ then the t-statistic is zero, its minimum value.
- In contrast, the KS-statistic is
  sensitive to any differences in the sample distributions.

Sensitivity to specific alternatives

Now consider a second dataset for which

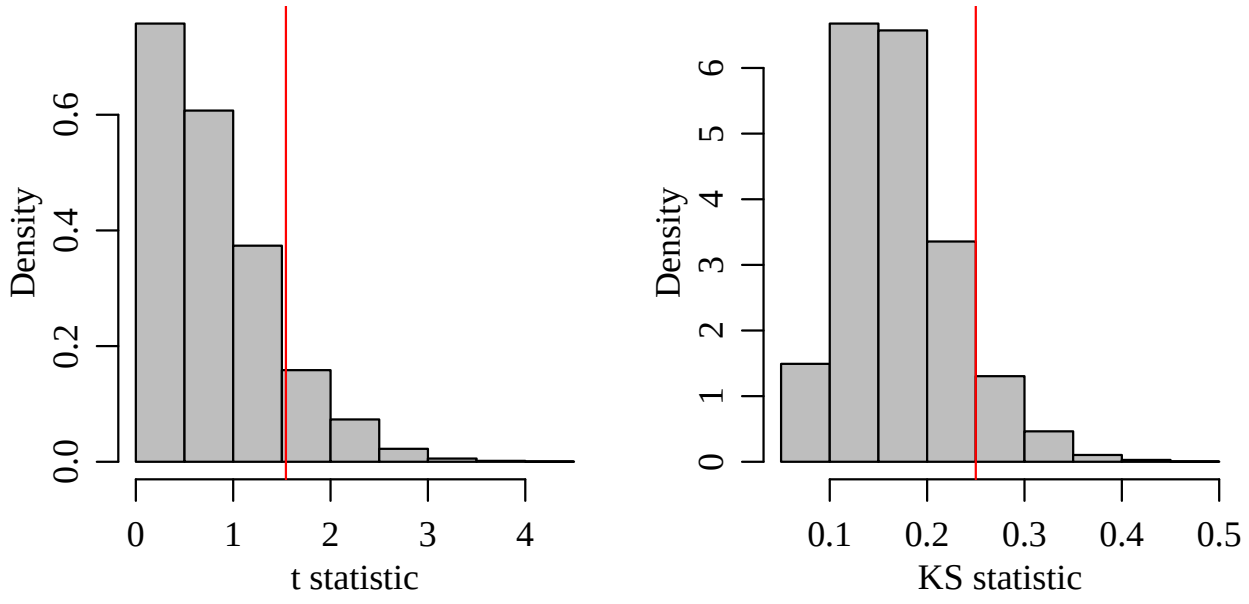
- $n_A = n_B = 40$
- $\bar{y}_A = 10.11$, $\bar{y}_B = 10.73$.
- $s_A = 1.75$, $s_B = 1.85$



The difference in sample means is about twice as large with the previous data. The sample standard deviations are pretty similar.

Is there evidence that the mean difference is caused by treatment?

Sensitivity to specific alternatives

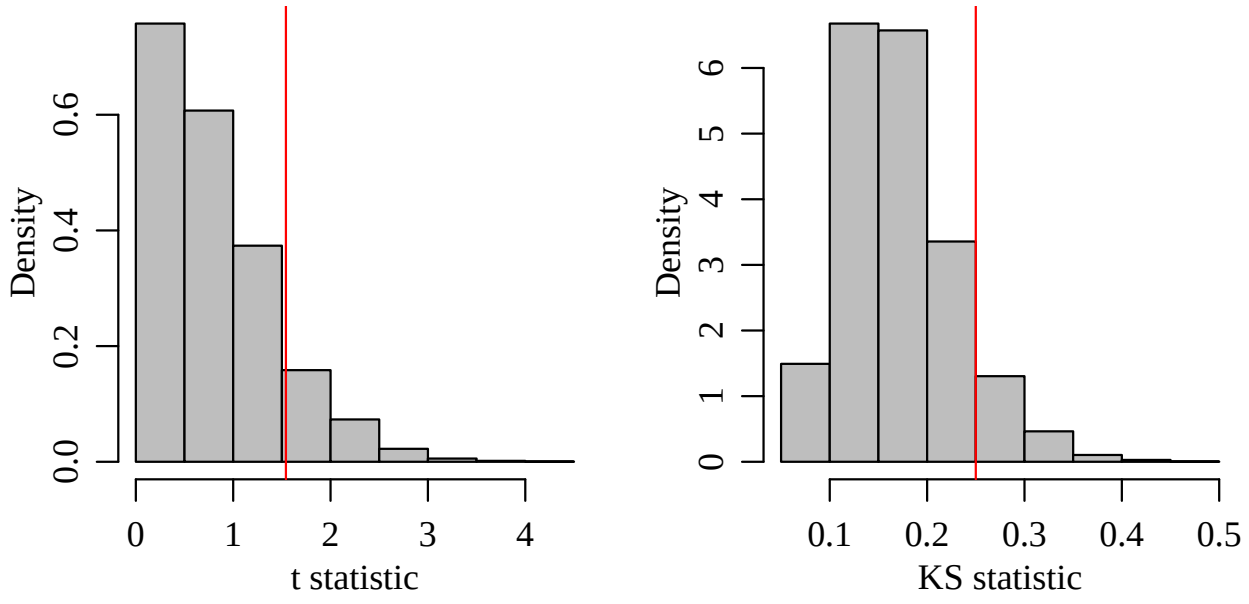


t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.54$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.54) = 0.122$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.25$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.25) = 0.106$

This time the two test statistics indicate similar evidence against H_0.
This is because the difference in the two sample distributions could primarily be summarized as the difference between the sample means, which the t-statistic can identify.

Sensitivity to specific alternatives

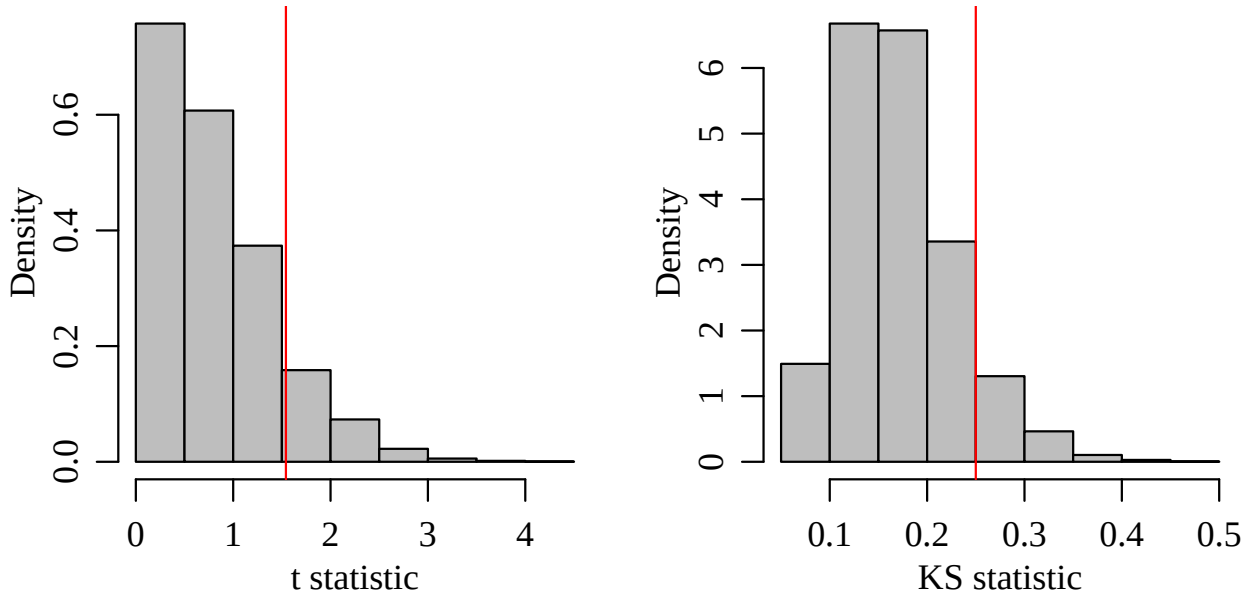


t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.54$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.54) = 0.122$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.25$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.25) = 0.106$

This time the two test statistics indicate similar evidence against H_0.
This is because the difference in the two sample distributions could primarily be summarized as the difference between the sample means, which the t-statistic can identify.

Sensitivity to specific alternatives

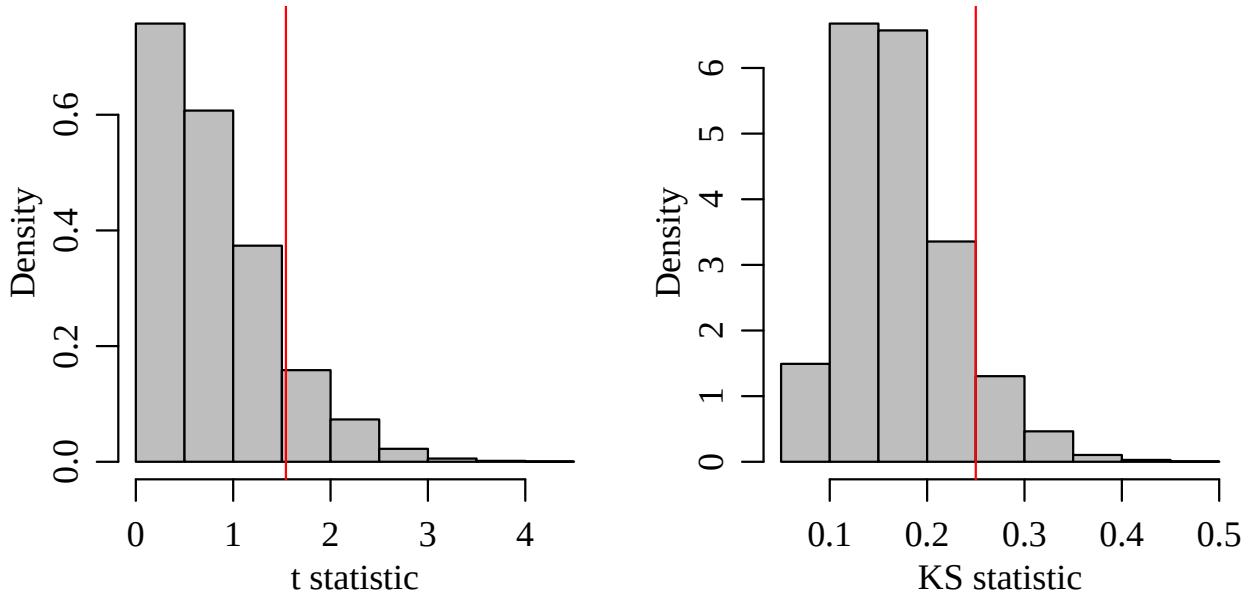


t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.54$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.54) = 0.122$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.25$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.25) = 0.106$

This time the two test statistics indicate similar evidence against H_0.
This is because the difference in the two sample distributions could primarily be summarized as the difference between the sample means, which the t-statistic can identify.

Sensitivity to specific alternatives



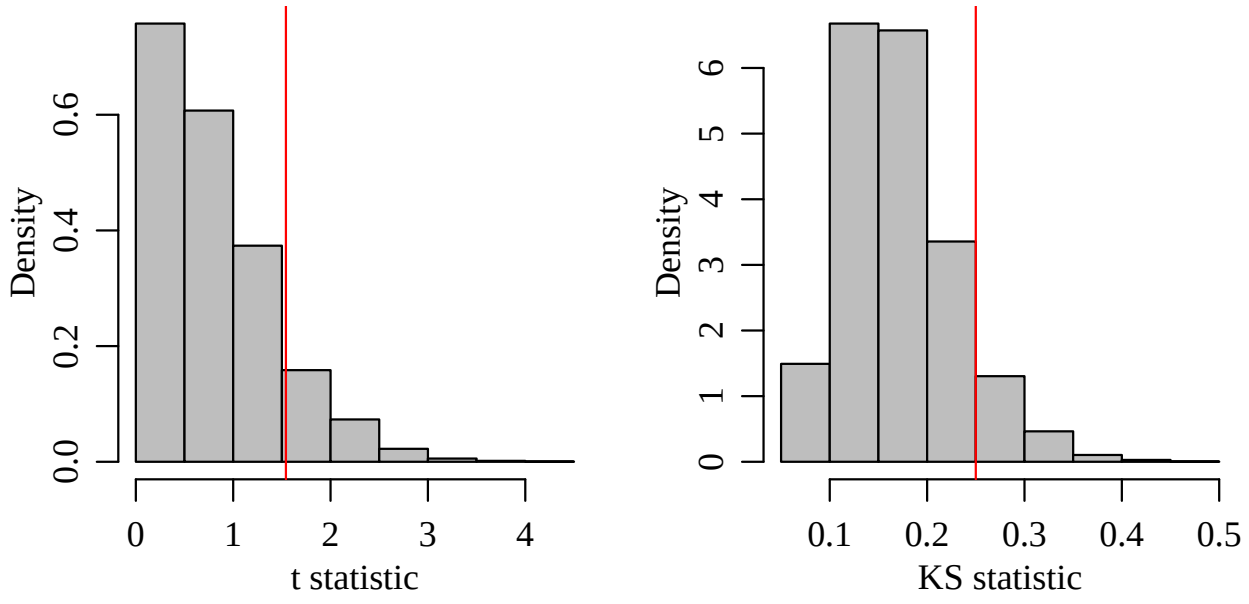
t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.54$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.54) = 0.122$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.25$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.25) = 0.106$

This time the two test statistics indicate similar evidence against H_0.

This is because the difference in the two sample distributions could primarily be summarized as the difference between the sample means, which the t-statistic can identify.

Sensitivity to specific alternatives



t-statistic : $g_t(\mathbf{y}_A, \mathbf{y}_B) = 1.54$, $\Pr(g_t(\mathbf{Y}_A, \mathbf{Y}_B) \geq 1.54) = 0.122$

KS-statistic: $g_{KS}(\mathbf{y}_A, \mathbf{y}_B) = 0.25$, $\Pr(g_{KS}(\mathbf{Y}_A, \mathbf{Y}_B) \geq 0.25) = 0.106$

This time the two test statistics indicate similar evidence against H_0.
This is because the difference in the two sample distributions could primarily be summarized as the difference between the sample means, which the t-statistic can identify.

Discussion

These last two examples suggest we should abandon g_t in favor of g_{KS} if we are interested in comparing the following hypothesis:

$$\begin{aligned} \mathrm{H}_0 &: \quad \text{treatment does not affect response} \\ \mathrm{H}_1 &: \quad \text{treatment does affect response} \end{aligned}$$

This is because, as we found, g_t is not sensitive all violations of H_0, it is only sensitive to violations of H_0 where there is a difference in means. However, in many situations we are actually interested in comparing the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: \quad \text{treatment does not affect response} \\ \mathrm{H}_1 &: \quad \text{treatment increases responses or decreases responses} \end{aligned}$$

In this case H_0 and H_1 are not complementary, and we are only interested in evidence against H_0 of a certain type, i.e. evidence that is consistent with H_1. In this situation we may want to use a statistic like g_t.

Discussion

These last two examples suggest we should abandon g_t in favor of g_{KS} if we are interested in comparing the following hypothesis:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment does affect response} \end{aligned}$$

This is because, as we found, g_t is not sensitive all violations of H_0, it is only sensitive to violations of H_0 where there is a difference in means. However, in many situations we are actually interested in comparing the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment increases responses or decreases responses} \end{aligned}$$

In this case H_0 and H_1 are not complementary, and we are only interested in evidence against H_0 of a certain type, i.e. evidence that is consistent with H_1. In this situation we may want to use a statistic like g_t.

Discussion

These last two examples suggest we should abandon g_t in favor of g_{KS} if we are interested in comparing the following hypothesis:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment does affect response} \end{aligned}$$

This is because, as we found, g_t is not sensitive all violations of H_0, it is only sensitive to violations of H_0 where there is a difference in means. However, in many situations we are actually interested in comparing the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment increases responses or decreases responses} \end{aligned}$$

In this case H_0 and H_1 are not complementary, and we are only interested in evidence against H_0 of a certain type, i.e. evidence that is consistent with H_1. In this situation we may want to use a statistic like g_t.

Discussion

These last two examples suggest we should abandon g_t in favor of g_{KS} if we are interested in comparing the following hypothesis:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment does affect response} \end{aligned}$$

This is because, as we found, g_t is not sensitive all violations of H_0, it is only sensitive to violations of H_0 where there is a difference in means. However, in many situations we are actually interested in comparing the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{treatment does not affect response} \\ \mathrm{H}_1 : &\quad \text{treatment increases responses or decreases responses} \end{aligned}$$

In this case H_0 and H_1 are not complementary, and we are only interested in evidence against H_0 of a certain type, i.e. evidence that is consistent with H_1. In this situation we may want to use a statistic like g_t.

Basic decision theory

Task: Accept or reject H_0 based on data.

	truth	
action	H_0 true	H_0 false
accept H_0	correct decision	type II error
reject H_0	type I error	correct decision

Recall: p-value $\approx \Pr(\text{data} \,|\mathrm{H}_0)$ (roughly speaking)

- the p-value can measure evidence against H_0;
- the smaller the p-value, the more evidence against H_0.

Basic decision theory

Task: Accept or reject H_0 based on data.

	truth	
action	H_0 true	H_0 false
accept H_0	correct decision	type II error
reject H_0	type I error	correct decision

Recall: p-value $\approx \Pr(\text{data} \,|\mathrm{H}_0)$ (roughly speaking)

- the p-value can measure evidence against H_0;
- the smaller the p-value, the more evidence against H_0.

Basic decision theory

Task: Accept or reject H_0 based on data.

	truth	
action	H_0 true	H_0 false
accept H_0	correct decision	type II error
reject H_0	type I error	correct decision

Recall: p-value $\approx \Pr(\text{data} \,|\mathrm{H}_0)$ (roughly speaking)

- the p-value can measure evidence against H_0;
- the smaller the p-value, the more evidence against H_0.

Basic decision theory

Task: Accept or reject H_0 based on data.

	truth	
action	H_0 true	H_0 false
accept H_0	correct decision	type II error
reject H_0	type I error	correct decision

Recall: p-value $\approx \Pr(\text{data} \,|\mathrm{H}_0)$ (roughly speaking)

- the p-value can measure evidence against H_0;
- the smaller the p-value, the more evidence against H_0.

Basic decision theory

Task: Accept or reject H_0 based on data.

	truth	
action	H_0 true	H_0 false
accept H_0	correct decision	type II error
reject H_0	type I error	correct decision

Recall: p-value $\approx \Pr(\text{data} \,|\mathrm{H}_0)$ (roughly speaking)

- the p-value can measure evidence against H_0;
- the smaller the p-value, the more evidence against H_0.

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error}|\mathrm{H}_0) \quad =$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error}|\mathrm{H}_0) \quad =$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.

It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error}|\mathrm{H}_0) \quad =$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error}|\mathrm{H}_0) \quad =$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error}|\mathrm{H}_0) \quad =$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error} | \mathrm{H}_0) \quad =$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\Pr(\text{type I error} | \mathrm{H}_0) \quad = \quad \Pr(\text{reject } \mathrm{H}_0 \, | \mathrm{H}_0)$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\begin{aligned} \Pr(\text{type I error}|\mathrm{H}_0) &= \Pr(\text{reject } \mathrm{H}_0 \,|\mathrm{H}_0) \\ &= \Pr(p\text{-value} \leq \alpha|\mathrm{H}_0) \end{aligned}$$

Decision procedure

1. Compute the *p*-value, comparing observed test statistic to null distribution.
2. Reject H_0 if the *p*-value $\leq \alpha$, otherwise accept H_0.

This procedure is called a **level-α test**.
It controls the pre-experimental probability of a **type I error**,
or for a series of experiments, controls the **type I error rate**.

$$\begin{aligned} \Pr(\text{type I error}|\mathrm{H}_0) &= \Pr(\text{reject } \mathrm{H}_0 \,|\mathrm{H}_0) \\ &= \Pr(p\text{-value} \leq \alpha|\mathrm{H}_0) \\ &= \alpha \end{aligned}$$

Interpretations of level-α tests

Single Experiment Interpretation: If you use a level-α test for your experiment where H_0 is true, then before you run the experiment there is probability α that you will erroneously reject H_0.

Many Experiments Interpretation: If level-α tests are used in a large population of experiments, then H_0 will be declared false in $(100 \times \alpha)\%$ of the experiments in which H_0 is true.

$$\begin{aligned} \Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true}) &= \alpha \\ \Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ true}) &= 1 - \alpha \end{aligned}$$

We need to be more specific than "H_0 false" in order to calculate the power. We need to specify *how* it is false.

Interpretations of level-α tests

Single Experiment Interpretation: If you use a level-α test for your experiment where H_0 is true, then before you run the experiment there is probability α that you will erroneously reject H_0.

Many Experiments Interpretation: If level-α tests are used in a large population of experiments, then H_0 will be declared false in $(100 \times \alpha)\%$ of the experiments in which H_0 is true.

$$\begin{aligned} \Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true}) &= \alpha \\ \Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ true}) &= 1 - \alpha \end{aligned}$$

We need to be more specific than "H_0 false" in order to calculate the power.
We need to specify *how* it is false.

Interpretations of level-α tests

Single Experiment Interpretation: If you use a level-α test for your experiment where H_0 is true, then before you run the experiment there is probability α that you will erroneously reject H_0.

Many Experiments Interpretation: If level-α tests are used in a large population of experiments, then H_0 will be declared false in $(100 \times \alpha)\%$ of the experiments in which H_0 is true.

$$\begin{aligned} \Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true}) &= \alpha \\ \Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ true}) &= 1 - \alpha \end{aligned}$$

We need to be more specific than "H_0 false" in order to calculate the power.
We need to specify *how* it is false.

Interpretations of level-α tests

Single Experiment Interpretation: If you use a level-α test for your experiment where H_0 is true, then before you run the experiment there is probability α that you will erroneously reject H_0.

Many Experiments Interpretation: If level-α tests are used in a large population of experiments, then H_0 will be declared false in $(100 \times \alpha)\%$ of the experiments in which H_0 is true.

$$\begin{aligned}
\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true}) &= \alpha \\
\Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ true}) &= 1 - \alpha \\
\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ false}) &= ? \\
\Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ false}) &= ?
\end{aligned}$$

$\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true})$ is the **level** and $\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ false})$ is the **power**.

We need to be more specific than "H_0 false" in order to calculate the power.
We need to specify *how* it is false.

Interpretations of level-α tests

Single Experiment Interpretation: If you use a level-α test for your experiment where H_0 is true, then before you run the experiment there is probability α that you will erroneously reject H_0.

Many Experiments Interpretation: If level-α tests are used in a large population of experiments, then H_0 will be declared false in $(100 \times \alpha)\%$ of the experiments in which H_0 is true.

$$\begin{aligned}
\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true}) &= \alpha \\
\Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ true}) &= 1 - \alpha \\
\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ false}) &= ? \\
\Pr(\mathrm{H}_0 \text{ accepted} | \mathrm{H}_0 \text{ false}) &= ?
\end{aligned}$$

$\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ true})$ is the **level** and $\Pr(\mathrm{H}_0 \text{ rejected} | \mathrm{H}_0 \text{ false})$ is the **power**.

We need to be more specific than "H_0 false" in order to calculate the power. We need to specify *how* it is false.