Comparing two samples

Eli Gurarie

QERM 598 - Lecture 2
University of Washington - Seattle

January 17, 2008

Ants

Seed ant (*Pogonomyrmex salinus*)

Thatch ant (*Formica planipilis*)

The Question:

WHICH IS BIGGER?

Seed ant (*Pogonomyrmex salinus*)

Thatch ant (*Formica planipilis*)

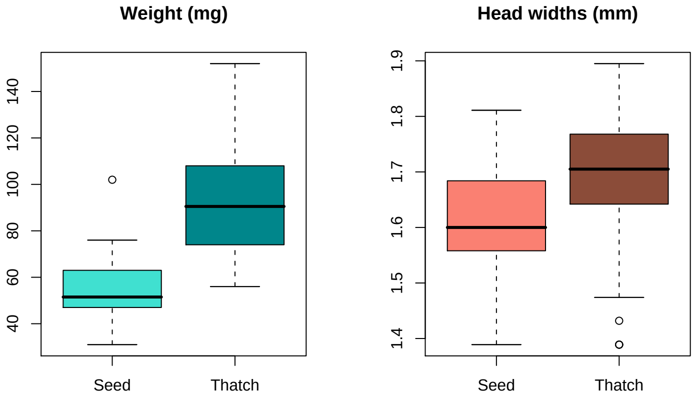
Step 1: Collect Data

	Seed Ant		Thatch Ant	
	Weight (mg)	Headwidth (mm)	Weight (mg)	Headwidth (mm)
1	51	1.600	90	1.642
2	55	1.726	104	1.895
3	53	1.558	106	1.684
4	48	1.474	57	1.432
5	31	1.389	90	1.811
6	72	1.642	132	1.684
7	45	1.558	91	1.768
8	65	1.684	110	1.768
9	50	1.600	86	1.726
10	102	1.811	152	1.895
11	57	1.684	74	1.600
12	38	1.642	58	1.389
13	67	1.600	71	1.389
14	57	1.558	79	1.642
15	76	1.811	67	1.474
16	67	1.684	112	1.853
17	43	1.558	103	1.726
18	50	1.600	61	1.726
19	35	1.516	141	1.768
20	65	1.642	81	1.642
21	41	1.600	103	1.726
22	63	1.768	56	1.474
23	48	1.726	81	1.642
24	59	1.558	91	1.642
25	44	1.558	130	1.768
26	60	1.516	59	1.389
27	48	1.600	91	1.726
28	52	1.726	108	1.811
29	51	1.600	125	1.811
30	47	1.642	75	1.642

note: data gratefully stolen from http://www.stat.ucla.edu/datasets/

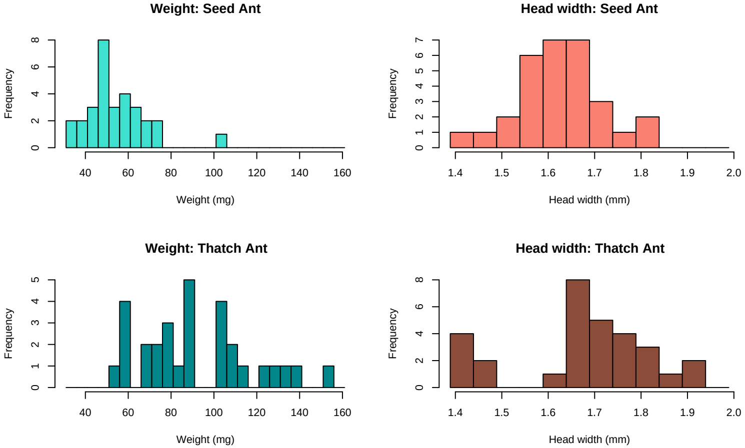
Step 2: Visualize Data

Boxplots!



Step 2: Visualize Data

Histograms!



Step 3: Summary statistics

		Seed Ant		Thatch Ant	
	N	$\bar{X}$	S	$\bar{X}$	S
Head width (mm)	30	14	195	92.8	26
Weight (mg)	30	1.62	0.096	1.67	0.147

A tough question

- It certainly *seems* like Thatch ants *might* to be bigger than Seed ants.
- But there are obviously *some* Seed ants that are bigger than *some* Thatch ants!
- What does the question "Which is Bigger?" actually mean?

The short answer

- It doesn't mean anything.

A tough question

- It certainly *seems* like Thatch ants *might* to be bigger than Seed ants.
- But there are obviously *some* Seed ants that are bigger than *some* Thatch ants!
- What does the question "Which is Bigger?" actually mean?

The short answer

- It doesn't mean anything.

A tough question

- It certainly *seems* like Thatch ants *might* to be bigger than Seed ants.
- But there are obviously *some* Seed ants that are bigger than *some* Thatch ants!
- What does the question "Which is Bigger?" actually mean?

The short answer

- It doesn't mean anything.

A tough question

- It certainly *seems* like Thatch ants *might* to be bigger than Seed ants.
- But there are obviously *some* Seed ants that are bigger than *some* Thatch ants!
- What does the question "Which is Bigger?" actually mean?

The short answer

- It doesn't mean anything.

A tough question

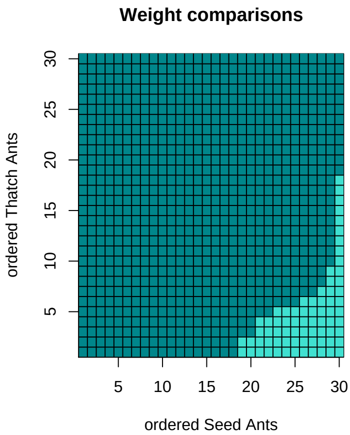
- It certainly *seems* like Thatch ants *might* to be bigger than Seed ants.
- But there are obviously *some* Seed ants that are bigger than *some* Thatch ants!
- What does the question "Which is Bigger?" actually mean?

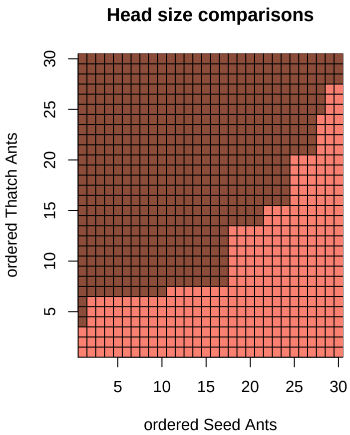
The short answer

- It doesn't mean anything.

We must refine the question...

E.g: what is the probability that *any given* Thatch Ant is bigger than *any given* Seed Ant?

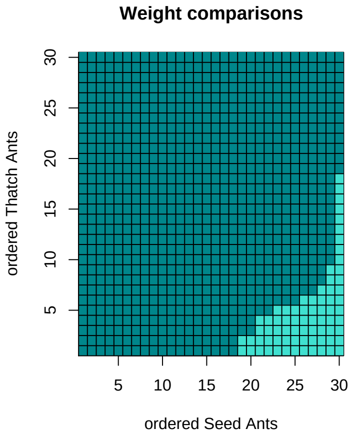


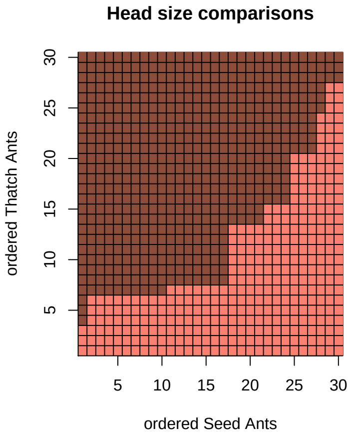


We must refine the question...

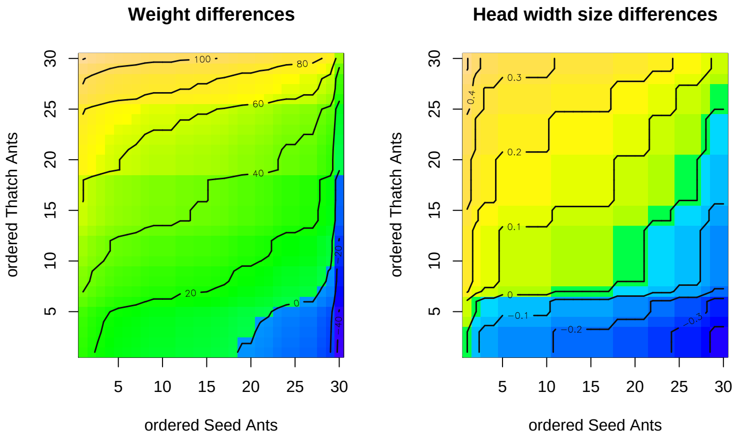
$$C_w = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{j=1}^{N_s} I(Wt_i > Ws_j) = \frac{872}{900} = 0.92$$

$$C_h = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{i=1}^{N_s} I(Ht_i > Hs_j) = \frac{559}{900} = 0.62$$





Or we can just go way too fancy ...



Getting there, but more questions!

- The statement that **A** is bigger than **B** *about* **X** % of the time is an improvement ...
- But how do we know that this comparison isn't an artifact of random sampling?

The short answer

- There is no short answer. It takes lots of really confusing statsy jargon to say anything about anything. Start getting used to it.

Getting there, but more questions!

- The statement that **A** is bigger than **B** *about* **X** % of the time is an improvement ...
- But how do we know that this comparison isn't an artifact of random sampling?

The short answer

- There is no short answer. It takes lots of really confusing statsy jargon to say anything about anything. Start getting used to it.

Getting there, but more questions!

- The statement that **A** is bigger than **B** *about* **X** % of the time is an improvement ...
- But how do we know that this comparison isn't an artifact of random sampling?

The short answer

- There is no short answer. It takes lots of really confusing statsy jargon to say anything about anything. Start getting used to it.

Getting there, but more questions!

- The statement that **A** is bigger than **B** *about* **X** % of the time is an improvement ...
- But how do we know that this comparison isn't an artifact of random sampling?

The short answer

- There is no short answer. It takes lots of really confusing statsy jargon to say anything about anything. Start getting used to it.

Getting there, but more questions!

- The statement that **A** is bigger than **B** *about* **X** % of the time is an improvement ...
- But how do we know that this comparison isn't an artifact of random sampling?

The short answer

- There is no short answer. It takes lots of really confusing statsy jargon to say anything about anything. Start getting used to it.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
   1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
   2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
   1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
   2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
    1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
    2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
    1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
    2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
   1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
   2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
   1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
   2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
   1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
   2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Hypothesis testing in 7 (or 6? or 8?) easy (ha! ha!) steps.

1. Since it can be tricky to even define what it is we want to know, we define *it's opposite*, which is often simpler. This is called the **null hypothesis**(H_0).
2. What the **null hypothesis** *isn't* we call the **alternative hypothesis** (H_1 or H_A).
3. We choose some summary of the data called the **test statistic** ($T_0 \sim f(t)$).
4. We create a **null distibution** of the test statistic... i.e. the distribution we would expect of the test statistic if the null hypothesis were true.
5. We calculate the experimental value of the **test statistic**, t_0, and compare it to our distribution.
6. We set some criterion, often called the **critical region**, within which we would *fail to reject* (not quite the same as "accept") the **null hypothesis**. Here, two things can happen:
    1. If $t_0$is "extreme" (lies outside our critical region), we reject the null hypothesis, accept the alternative hypothesis, humbly acknowledging that we *might* be wrong, and call the probability that we might be wrong the **Type I error**.
    2. If $t_0$is not "extreme", we *fail to reject* the null hypothesis, calling the probability that we *might* be wrong the **Type II error**.

Example: Step 1-2, Null and Alternative Hypotheses

$=$

H_0: Seed and Thatch ants can be considered to come from the "same" population.

$\neq$

H_1: Seed and Thatch ants come from different populations

Example: Step 3 - Choose test statistic

We *could* do something crazy, like the count statistic:

$$C_w = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{j=1}^{N_s} I(W_{ti} > W_{sj})$$
$$C_h = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{i=1}^{N_s} I(H_{ti} > H_{sj})$$

But that's kind of crazy. How about something relatively straightforward ... like the difference between the means?

$$t_W = \bar{W}_t - \bar{W}_s$$
$$t_H = \bar{H}_t - \bar{H}_s$$

Example: Step 3 - Choose test statistic

We *could* do something crazy, like the count statistic:

$$C_w = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{j=1}^{N_s} I(W_{ti} > W_{sj})$$
$$C_h = \frac{1}{N_t N_s} \sum_{i=1}^{N_t} \sum_{i=1}^{N_s} I(H_{ti} > H_{sj})$$

But that's kind of crazy. How about something relatively straightforward ... like the difference between the means?

$$t_W = \bar{W}_t - \bar{W}_s$$
$$t_H = \bar{H}_t - \bar{H}_s$$

Example: Step 4 - Obtain null-distribution

One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

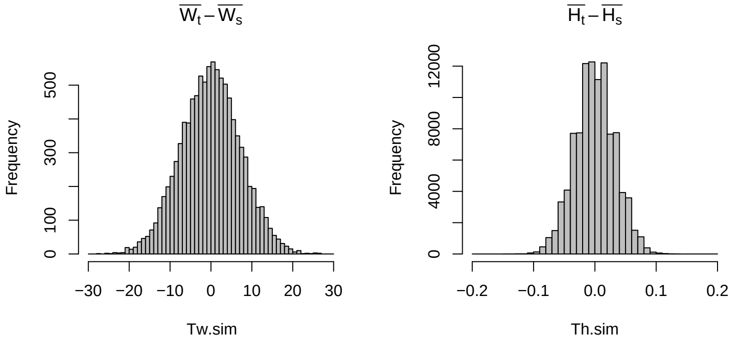
- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

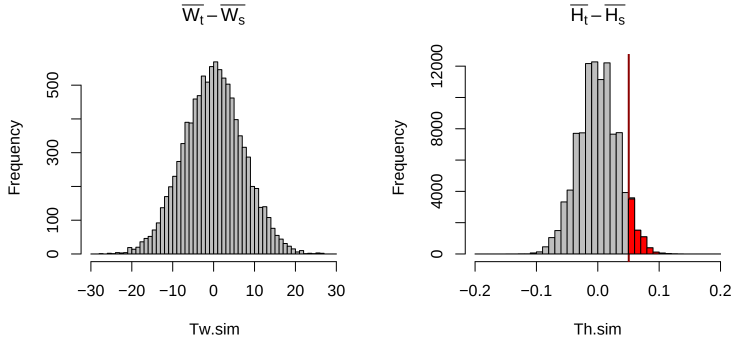
One approach is using **Monte Carlo simulation** to obtain a simulated null-distribution of the test statistic

- If the null hypothesis is true, then there is **no** difference between the two groups means we can resample them in any which way
- So ....
  1. shuffle all weights W
  2. split up into two new vectors: $W_{S.sim}$ and $W_{T.sim}$
  3. obtain and store the statistic $T_{W.sim} = \bar{W}_{T.sim} - \bar{W}_{T.sim}$
  4. Repeat steps 1-3 a bunch of times.
- Repeat steps 1-4 for head sizes H.

Example: Step 4 - Obtain null-distribution

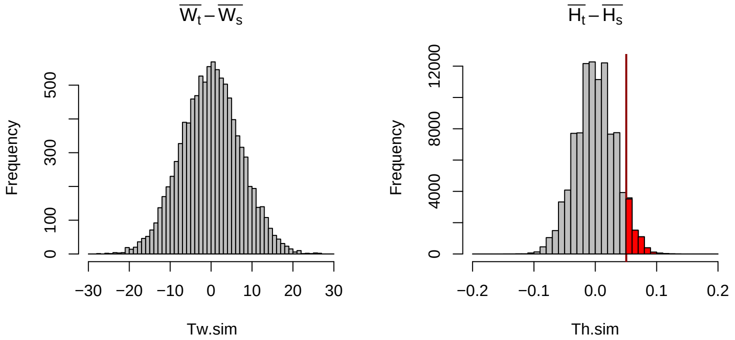


Example: Step 5 - Assess observed statistic



$t_w = 38.13333$ $\quad$ $t_h = 0.0504667$

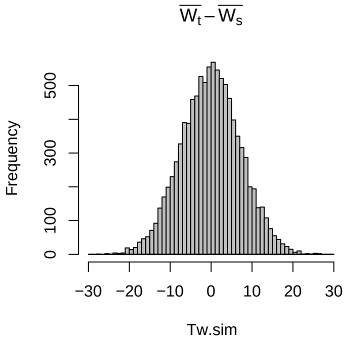
Example: Step 6a - is this extreme enough?

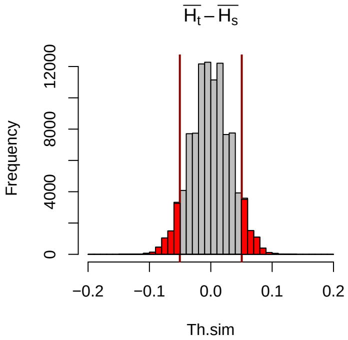


$\Pr\{T_{W.sim} > T_W\} = 0$ $\Pr\{T_{H.sim} > T_H\} = 0.0598$

Example: Step 6b - is this extreme enough?

The measure of "extremeness" *shoulg* reflect the fact that H_1 is two-sided!





$\Pr\{T_{W.sim} > T_W\} = 0$

$\Pr\{|T_{H.sim}| > T_H\} = 0.1188$

Example: Monte Carlo simulation - Step 7

Decide: is this extreme enough? (aka the hoodoo-voodoo step)

- After 10,000 simulations of random samplings of Weight under the null hypothesis, there were exactly 0 whose mean difference was more extreme than our measured difference of 38.1 mg. Thus we can **reject the null hypothesis with high confidence**.
- After 10,000 simulations of random samplings of Head size under the null hypothesis, about 11% had values that more extreme than the measured difference in means of 0.051 mm... We *could* still "reject the null hypothesis", but not with very high confidence since there's a 1 in 10 chance that a sampling from the null hypothesis will yield a more extreme result than our data. A "typical" **significance level** is 0.05, but this is partially a historical artifact from the days when everyone relied on tables. If we finagled our hypotheses to be one-sided ($H_0 : H_t \leq H_s$ $H_t > H_s$), the p-value drops to 0.059. Is that good enough? It's not *strictly* below 0.05. It's the sort of result that might be classified as "marginally significant".
- Let's say we really feel that it's not extreme enough. Does that mean the null hypothesis is true? NO! It just means we lacked to power to reject it. We really, really wanted to, but we **failed to reject** H_0.
- **See why we call this the hoodoo-voodoo step?**

Example: Monte Carlo simulation - Step 7

Decide: is this extreme enough? (aka the hoodoo-voodoo step)

- After 10,000 simulations of random samplings of Weight under the null hypothesis, there were exactly 0 whose mean difference was more extreme than our measured difference of 38.1 mg. Thus we can **reject the null hypothesis with high confidence**.
- After 10,000 simulations of random samplings of Head size under the null hypothesis, about 11% had values that more extreme than the measured difference in means of 0.051 mm... We *could* still "reject the null hypothesis", but not with very high confidence since there's a 1 in 10 chance that a sampling from the null hypothesis will yield a more extreme result than our data. A "typical" **significance level** is 0.05, but this is partially a historical artifact from the days when everyone relied on tables. If we finagled our hypotheses to be one-sided ($H_0 : H_t \leq H_s \; H_t > H_s$), the p-value drops to 0.059. Is that good enough? It's not *strictly* below 0.05. It's the sort of result that might be classified as "marginally significant".
- Let's say we really feel that it's not extreme enough. Does that mean the null hypothesis is true? NO! It just means we lacked to power to reject it. We really, really wanted to, but we **failed to reject** H_0.
- **See why we call this the hoodoo-voodoo step?**

Example: Monte Carlo simulation - Step 7

Decide: is this extreme enough? (aka the hoodoo-voodoo step)

- After 10,000 simulations of random samplings of Weight under the null hypothesis, there were exactly 0 whose mean difference was more extreme than our measured difference of 38.1 mg. Thus we can **reject the null hypothesis with high confidence**.
- After 10,000 simulations of random samplings of Head size under the null hypothesis, about 11% had values that more extreme than the measured difference in means of 0.051 mm... We *could* still "reject the null hypothesis", but not with very high confidence since there's a 1 in 10 chance that a sampling from the null hypothesis will yield a more extreme result than our data. A "typical" **significance level** is 0.05, but this is partially a historical artifact from the days when everyone relied on tables. If we finagled our hypotheses to be one-sided ($H_0 : H_t \leq H_s \; H_t > H_s$), the p-value drops to 0.059. Is that good enough? It's not *strictly* below 0.05. It's the sort of result that might be classified as "marginally significant".
- Let's say we really feel that it's not extreme enough. Does that mean the null hypothesis is true? NO! It just means we lacked to power to reject it. We really, really wanted to, but we **failed to reject** H_0.
- See why we call this the hoodoo-voodoo step?

Example: Monte Carlo simulation - Step 7

Decide: is this extreme enough? (aka the hoodoo-voodoo step)

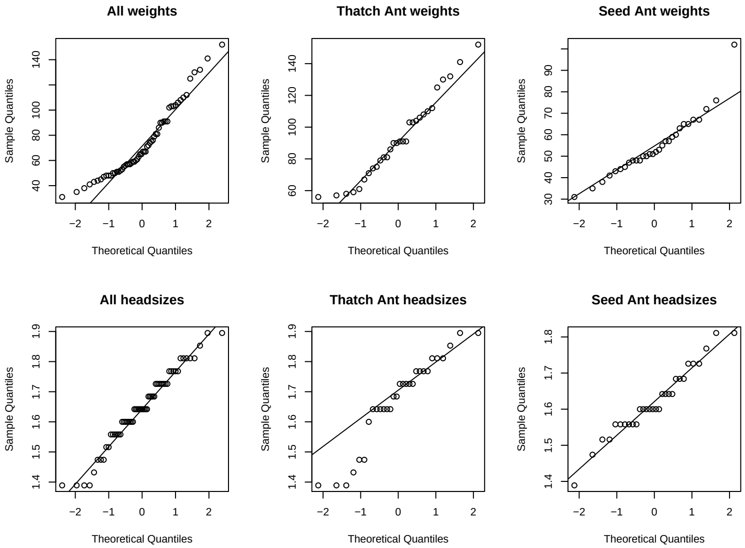
- After 10,000 simulations of random samplings of Weight under the null hypothesis, there were exactly 0 whose mean difference was more extreme than our measured difference of 38.1 mg. Thus we can **reject the null hypothesis with high confidence**.
- After 10,000 simulations of random samplings of Head size under the null hypothesis, about 11% had values that more extreme than the measured difference in means of 0.051 mm... We *could* still "reject the null hypothesis", but not with very high confidence since there's a 1 in 10 chance that a sampling from the null hypothesis will yield a more extreme result than our data. A "typical" **significance level** is 0.05, but this is partially a historical artifact from the days when everyone relied on tables. If we finagled our hypotheses to be one-sided ($H_0 : H_t \leq H_s$ $H_t > H_s$), the p-value drops to 0.059. Is that good enough? It's not *strictly* below 0.05. It's the sort of result that might be classified as "marginally significant".
- Let's say we really feel that it's not extreme enough. Does that mean the null hypothesis is true? NO! It just means we lacked to power to reject it. We really, really wanted to, but we **failed to reject** H_0.
- **See why we call this the hoodoo-voodoo step?**

T-Tests

In statistics, lots and lots of magical things happen when you make a few assumptions. The biggies are:

- Independence
  - (also necessary for Monte Carle, Randomization, etc. etc.)
- Constant variance between groups that are being compared
- Normality

T-tests: Assess normality



T-tests: some basic math facts

- if $X_1, X_2, ...X_n$ are iid rv's with distribution $\mathrm{N}\left\{\mu, \sigma^2\right\}$ then:

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu} \sim \mathrm{N}\left\{\mu, \frac{\sigma^2}{n}\right\} \tag{1}$$

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \text{Chi-squared}\{n\} \tag{2}$$

- if $Y \sim \mathrm{N}\{0,1\}$ and $Z \sim \text{Chi-squared}\{n\}$ then:

$$\frac{Y}{\sqrt{Z/n}} \sim T\{n\} \tag{3}$$

  where $T\{n\}$ is Student's-T distribution with n degrees of freedom.

- Exercise: Combine all these facts to show that *under the assumption that* $\mu = 0$,

$$\frac{\sqrt{n(n-1)}\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim T\{n-1\} \tag{4}$$

T-tests: a little more math

- Consider n_1 measurements of $X_1 \sim \mathrm{N}\left\{\mu_1, \sigma_1^2\right\}$and n_2 measurements of $X_2 \sim \mathrm{N}\left\{\mu_2, \sigma_2^2\right\}$.
- Assume: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Under H_0: $\mu_1 = \mu_2$.
- With these conditions, we can derive the following result:

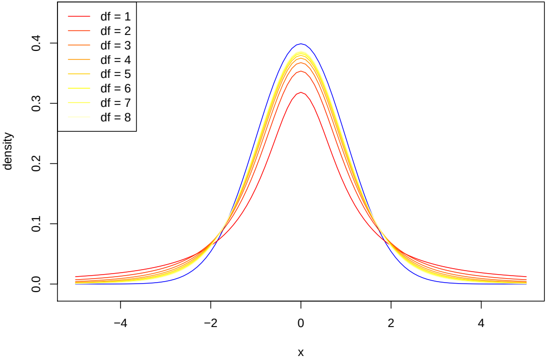
$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \tag{5}$$

where S_p is called the **pooled variance** and is a weighted estimate of σ^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} \tag{6}$$

T-tests: long story short

This beast: $t_0 = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ is a **test-statistic** with a known **null distribution** T_0 which looks a lot like the **standard normal distribution**

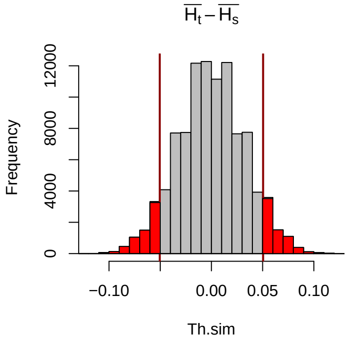


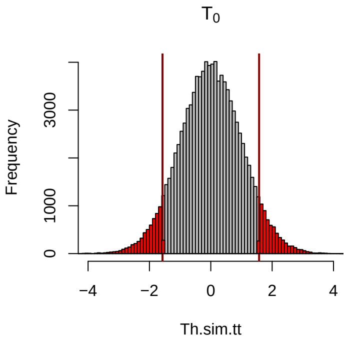
T-tests: Check our data

- For weight: $t_w = 7.0841$
  assess against T_{58}: $\Pr\{|T_{58}| > t_w\} = 10^{-9}$, so **REJECT** H_0

- For headsize: $t_h = 1.157$
  assess against t_{58}: $\Pr\{|T_{58}| > t_h\} = .1217$, so **FAIL TO REJECT** H_0

T-tests: Compare methods

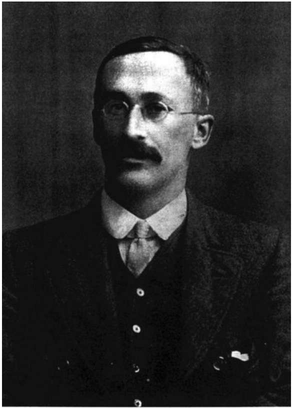




$\Pr\{T_{W.sim} > T_H\} = .1188$

$\Pr\{T_{58} > t_h\} = 0.1217$

Meet Mr. William Sealy Gosset



'Student' in 1908

William Gosset (June 13, 1876 - October 16, 1937) - the inventor of the T-test - was a bright mathematician who worked for the Guiness brewing company. Some time earlier, a scientist had inadvertently revealed important brewing secrets in a science journal, so the company decided to clamp down on publishing careers. Gosset did his statistics late at night and published pseudonymously as "Student" (hence Student's-T). He went through much work hand-checking estimates working on the small sample problem. Apparently the company was too stingy with it's wares for him to perform experiments on large samples. Or perhaps, he felt sorry for his liver.