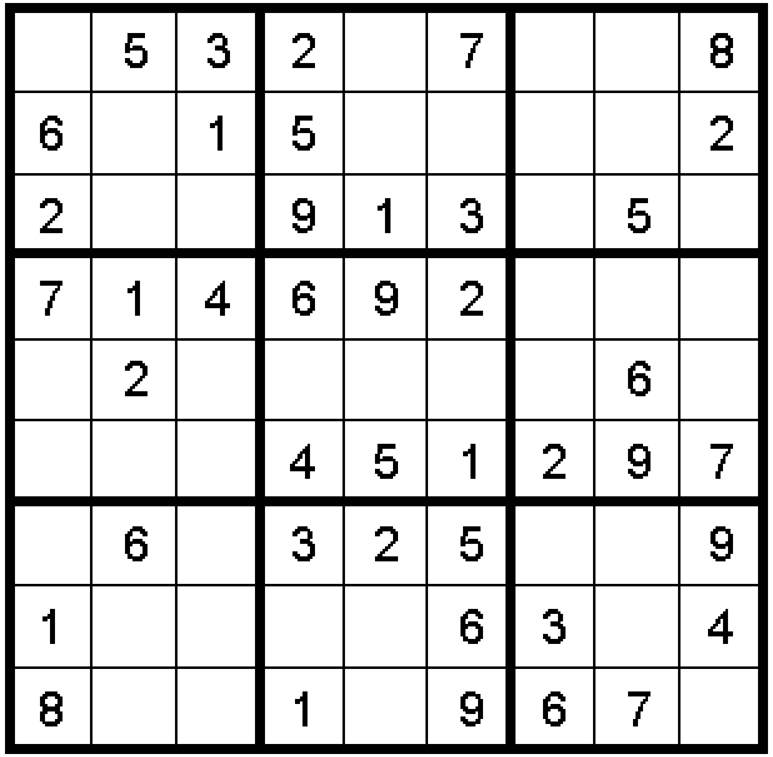
Hypothesis tests for more than two means and for correlation



	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Overview

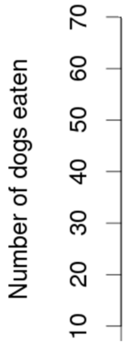
Taking stock of where we are and where we are going

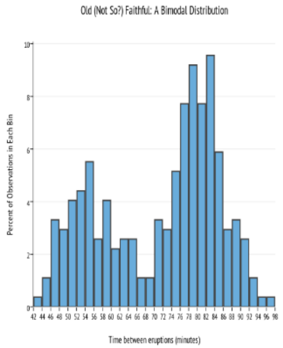
Hypothesis tests for more than two means continued

Hypothesis tests for correlation

Where we are: what we have covered

Descriptive statistics

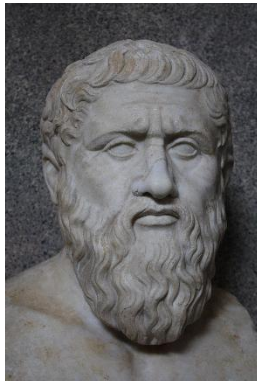


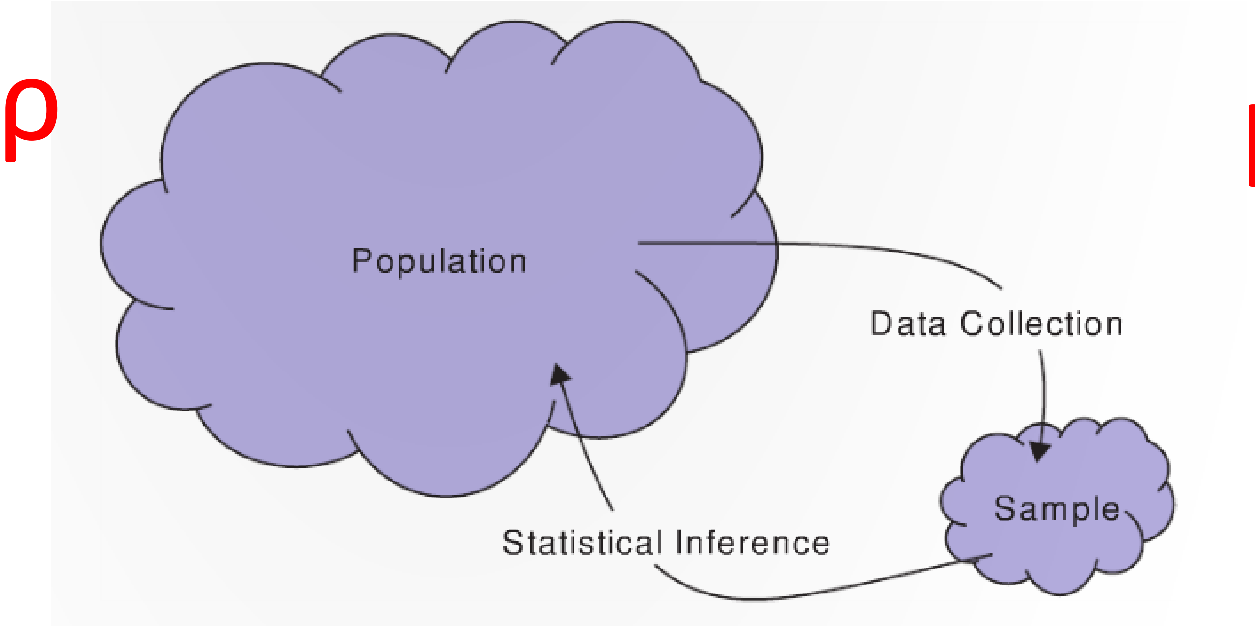


Statistical inference

- We have used computational methods for inference

π, μ, σ, ρ



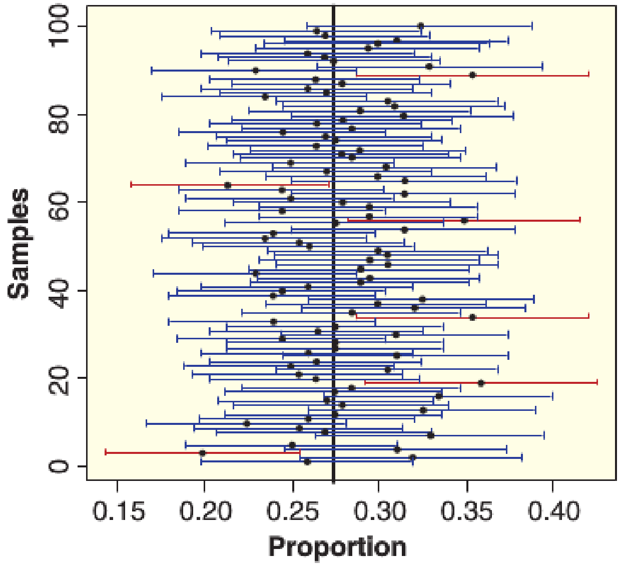


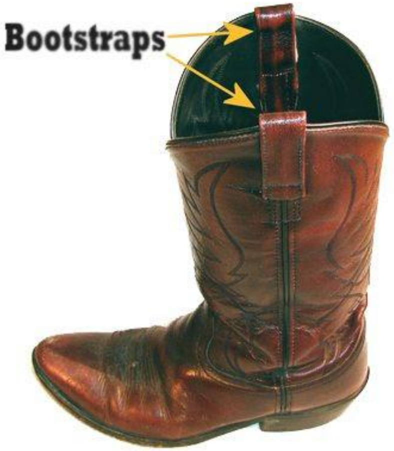
$\hat{p}, \bar{x}, s, r$

Where we are: what we have covered

Statistical inference using computational methods

- Confidence intervals using the bootstrap
- Hypothesis tests using permutation/randomization tests
  - Single proportion and two means





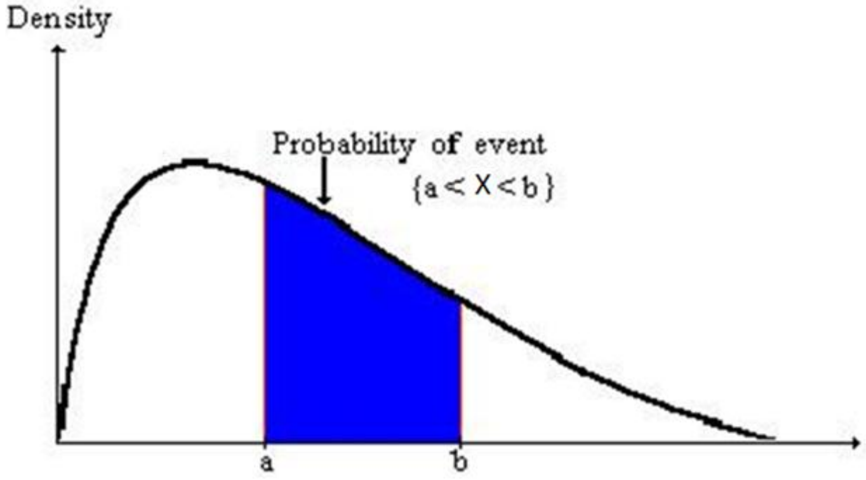
Where we are: where are we going…

Continuation of statistical inference using computational methods

- Hypothesis tests for more than two means and correlation
- Theories of hypothesis tests

Statistical inference based on math/theory

- t-tests, ANOVA, regression, etc.



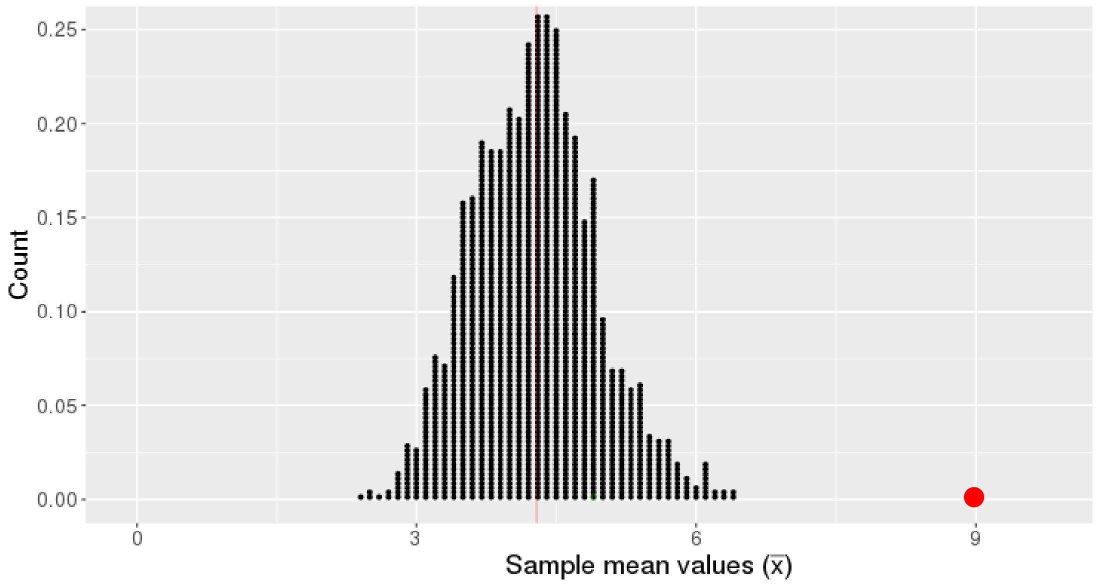
Hypothesis tests for more than two means continued

The logic of hypothesis tests…

We start with a claim about a population parameter

- E.g., μ = 4[illegible]

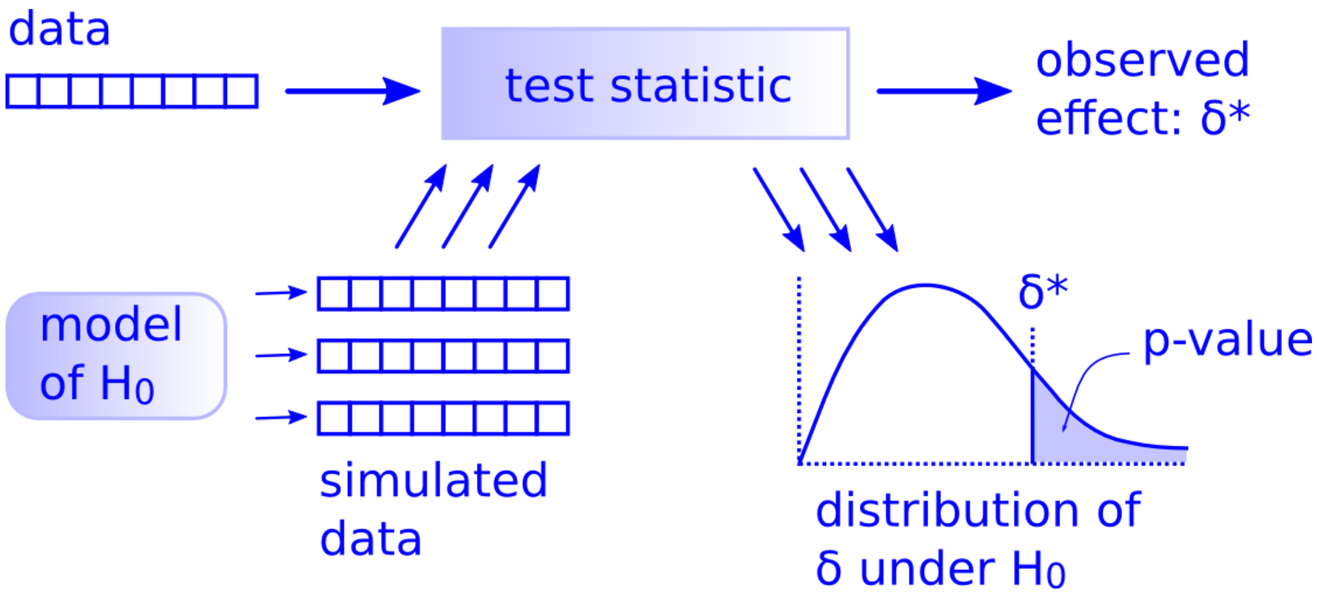
This claim implies we should get a certain distribution of statistics

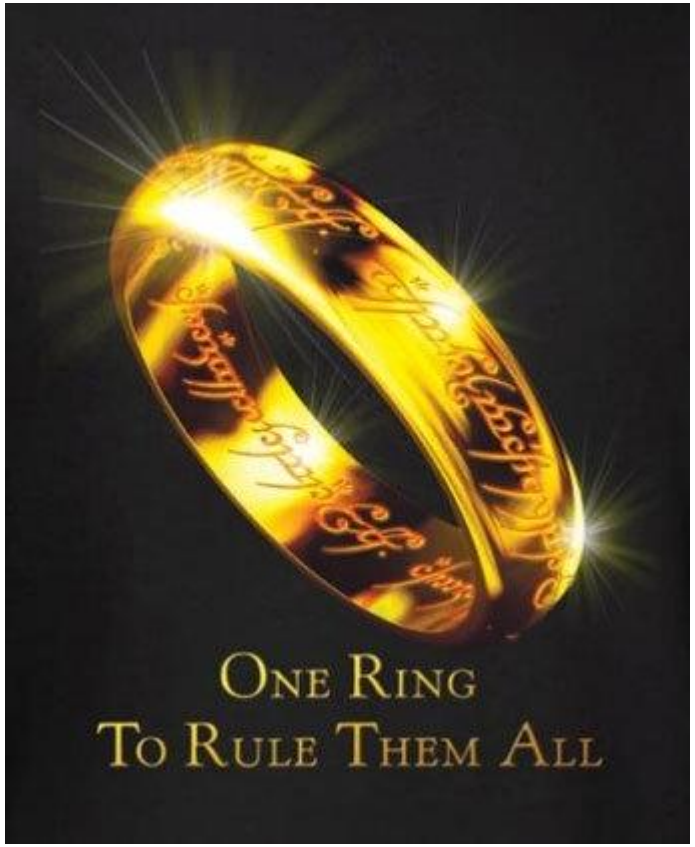


If our observed statistic is highly unlikely, we reject the claim

The logic of hypothesis tests…

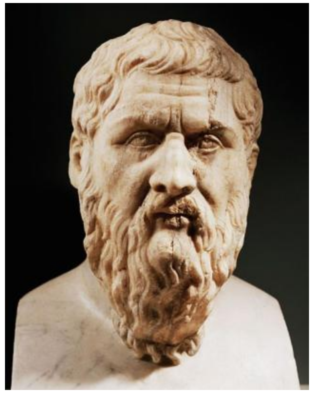
There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

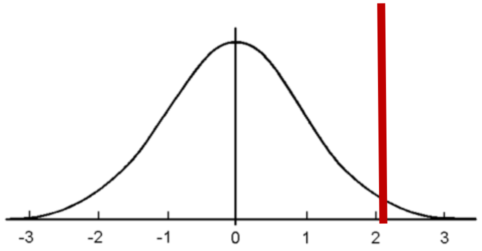
Five steps of hypothesis testing



1. State H_0 and H_A
    - Assume Gorgias (H_0) was right

2. Calculate the actual observed statistic

3. Create a **null distribution** of statistics that are consistent with H_0
    - i.e., a distribution of statistics that we would expect if Gorgias is right

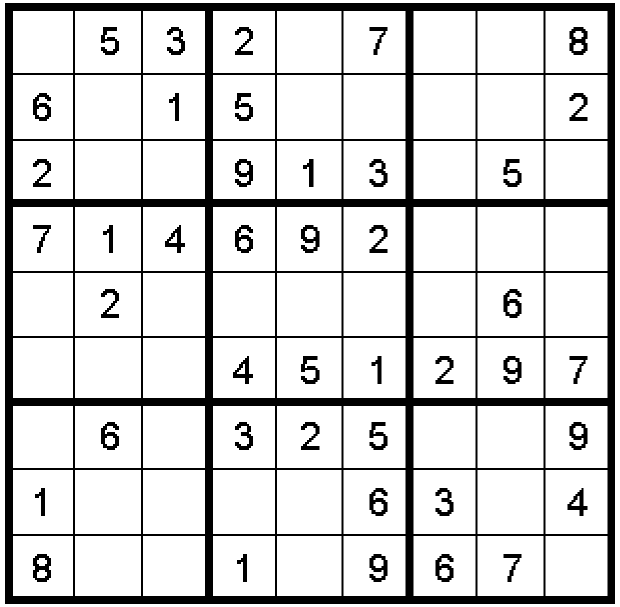


4. Get the probability we would get a statistic more than the observed statistic from the null distribution
    - p-value

5. Make a judgement
    - Assess whether the results are statistically significant

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student’s major and the time it takes to complete a small Sudoku-like puzzle



	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student’s major and the time it takes to complete a small Sudoku-like puzzle

They grouped majors into four categories

- Applied science (as)
- Natural science (ns)
- Social science (ss)
- Arts/humanities (ah)

What is the first step of hypothesis testing?

Sudoku by field

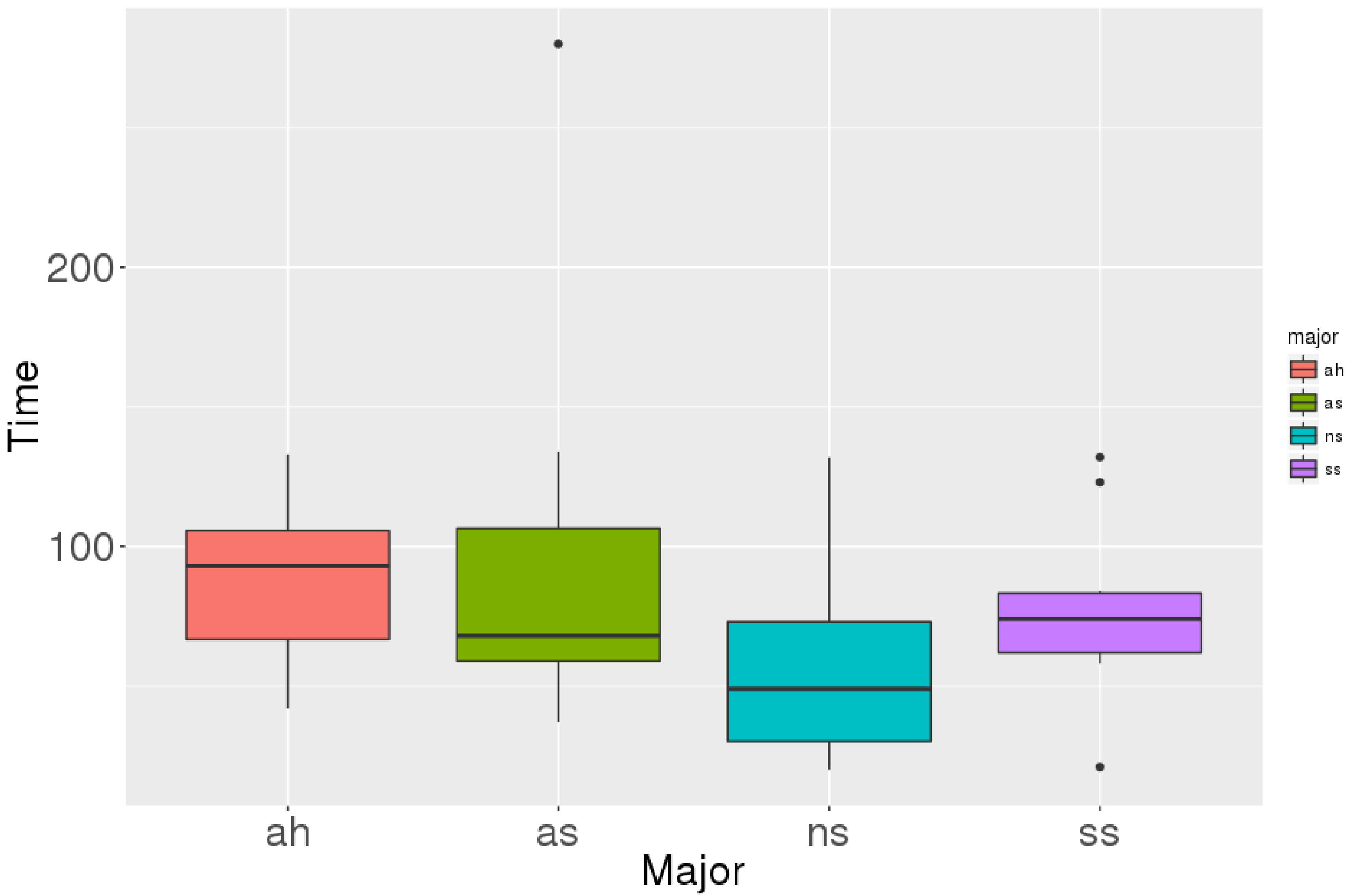
1. State the null and alternative hypotheses!

H_0: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$
H_A: $\mu_i \neq \mu_j$ for one pair of fields of study

What should we do next?

Let’s plot the data first…

Step 2a: Plot of completion time by major



What should we do next?

Sudoku by field

1. State the null and alternative hypotheses!

H$_0$: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$

H$_A$: $\mu_i \neq \mu_j$ for one pair of fields of study

Thoughts on the statistic of interest?

Comparing multiple means

There are many possible statistics we could use. A few choices are:

1. Group range statistic:

   max $\bar{x}$ - min $\bar{x}$

2. Mean absolute difference (MAD):

   $(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$

3. F statistic:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

Using the MAD statistic

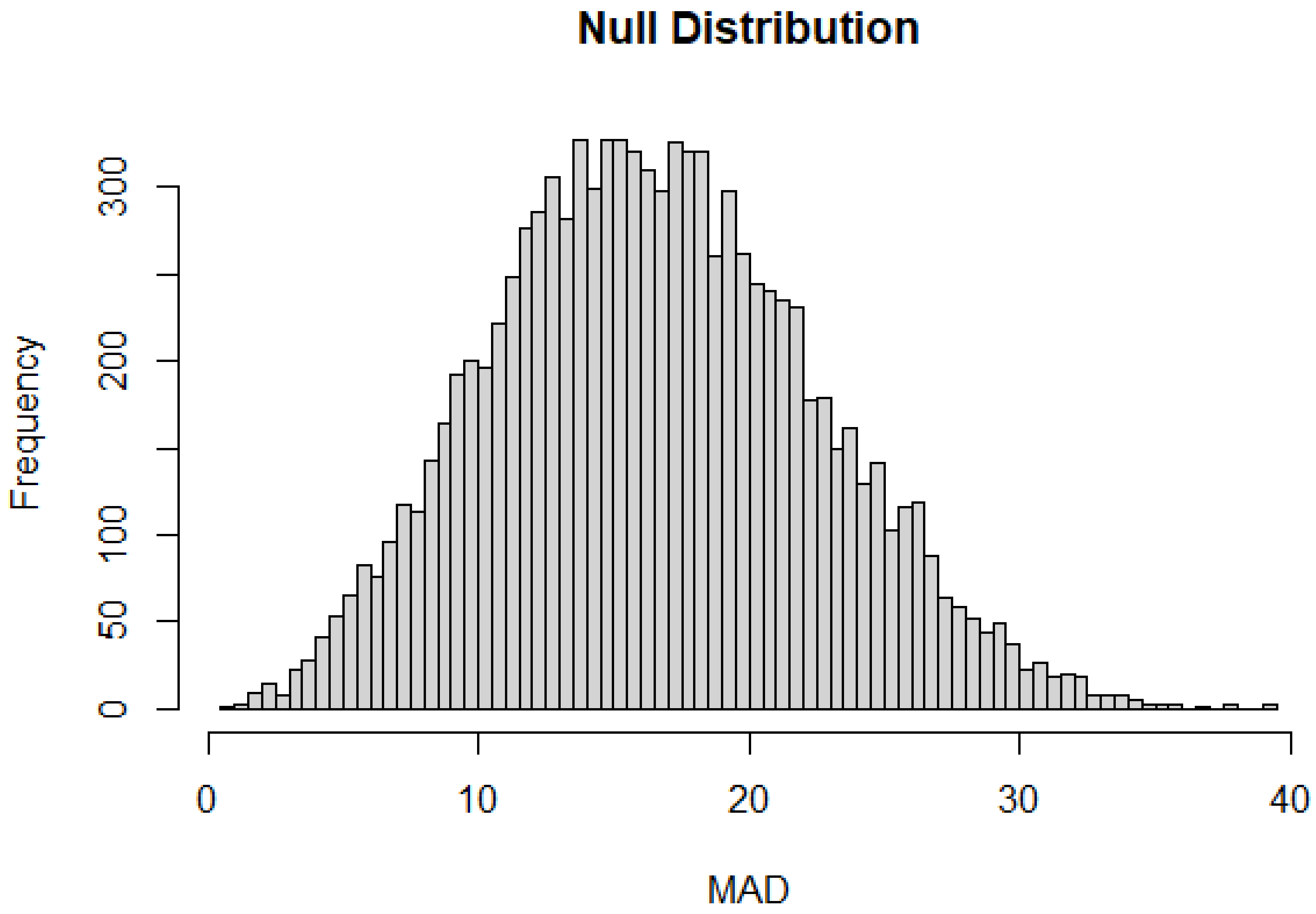
Mean absolute difference (MAD):

$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$

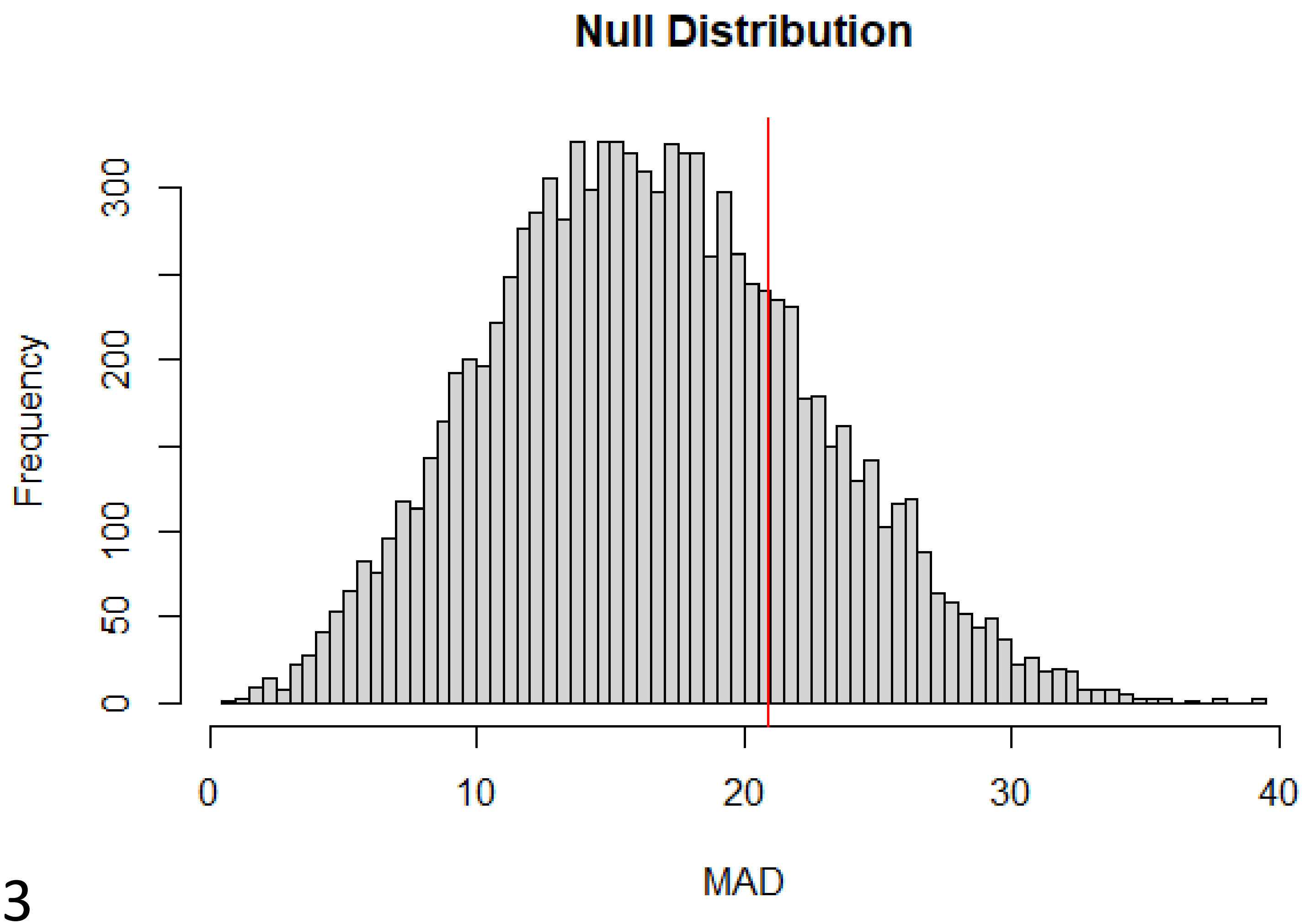
Observed statistic value = 20.88

How can we create the null distribution?

Null distribution



P-value



p-value = .233

Conclusions?

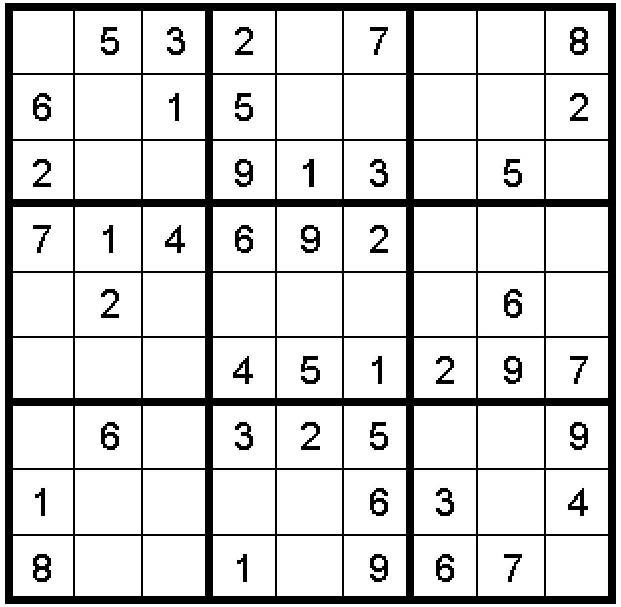
Hypothesis tests for more than two means in R

Hypothesis tests for more than two means in R

Step 1: null and alternative hypotheses…

H_0: $\mu_{as} = \mu_{ns} = \mu_{ss} = \mu_{ah}$

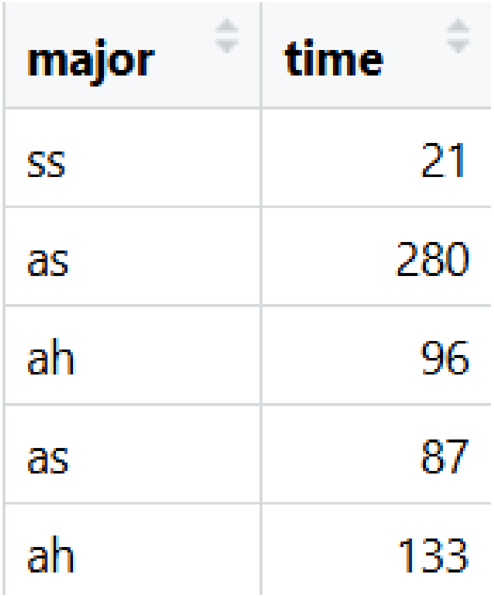
H_A: $\mu_i \neq \mu_j$ for one pair of fields of study



	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Let’s try this analysis in R…

```
# get the data
sudoku_data <- read.table("MajorPuzzle.txt", header = TRUE)

# Extract vectors from the data frame
completion_times <- sudoku_data$time
majors <- sudoku_data$major
```

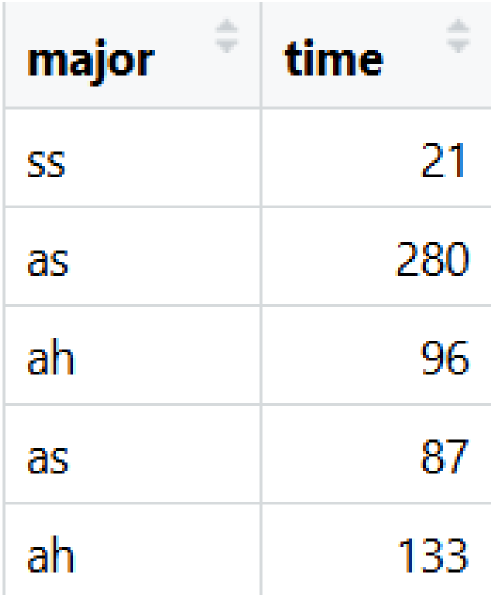


major	time
ss	21
as	280
ah	96
as	87
ah	133

Visualize the data

How can we visualize the data?

```
# We can create side-by-side boxplots using
boxplot(completion_time ~ major,
        xlab = "Major",   ylab = "Time (s)")
```



major	time
ss	21
as	280
ah	96
as	87
ah	133

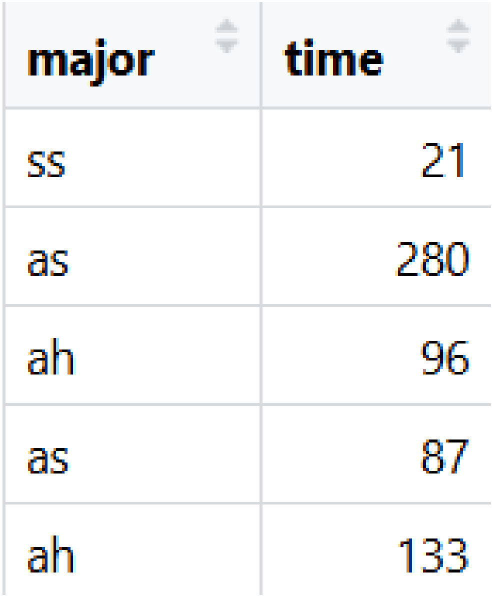
Calculating the statistic of interest

We can get the MAD statistic using the get_MAD_stat() function

$$\text{MAD} = (|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$$

get_MAD_stat(data_vector, grouping_vector)

- data_vector: a vector of quantitative data
- grouping_vector: a vector of categorical data indicating which group the quantitative data is in



major	time
ss	21
as	280
ah	96
as	87
ah	133

Can you get the MAD statistic for the sudoku data?

```
obs_stat <-  get_MAD_stat(completion_time, major)
```

Creating the null distribution

Q: How could we create one point in a null distribution?

- A: Shuffle the grouping_vector (major vector) and calculate the MAD statistic

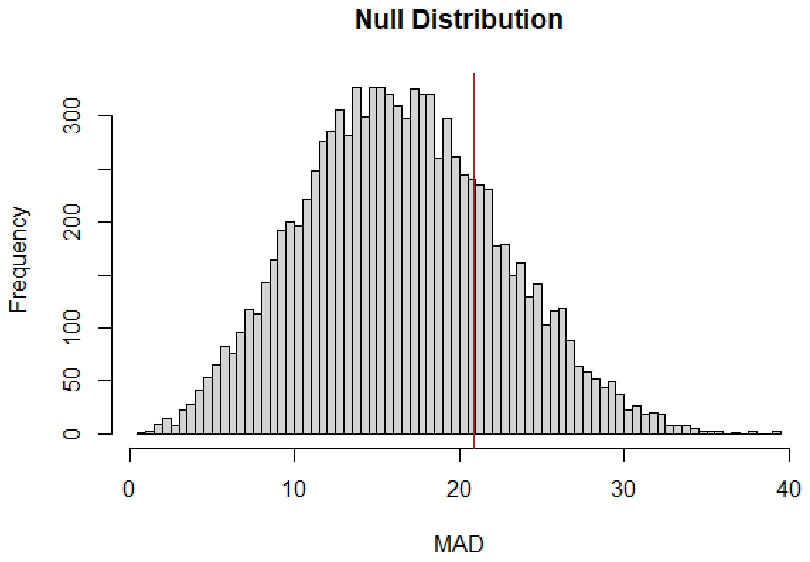
Q: How can we do this in R?

```
shuffled_majors <- shuffle(major)

get_MAD_stat(completion_time, shuffled_majors)
```

Creating the null distribution

Q: How can we create a full null distribution?

```
null_dist <- do_it(10000) * {
        shuffled_majors <- shuffle(majors)
        get_MAD_stat(completion_times, shuffled_majors)
}

# visualize the null distribution
hist(null_dist, breaks = 200)
abline(v = obs_stat, col = "red")
```



Steps 4 and 5

Q: What do we do next and how do we do it?

- A: We get the p-value

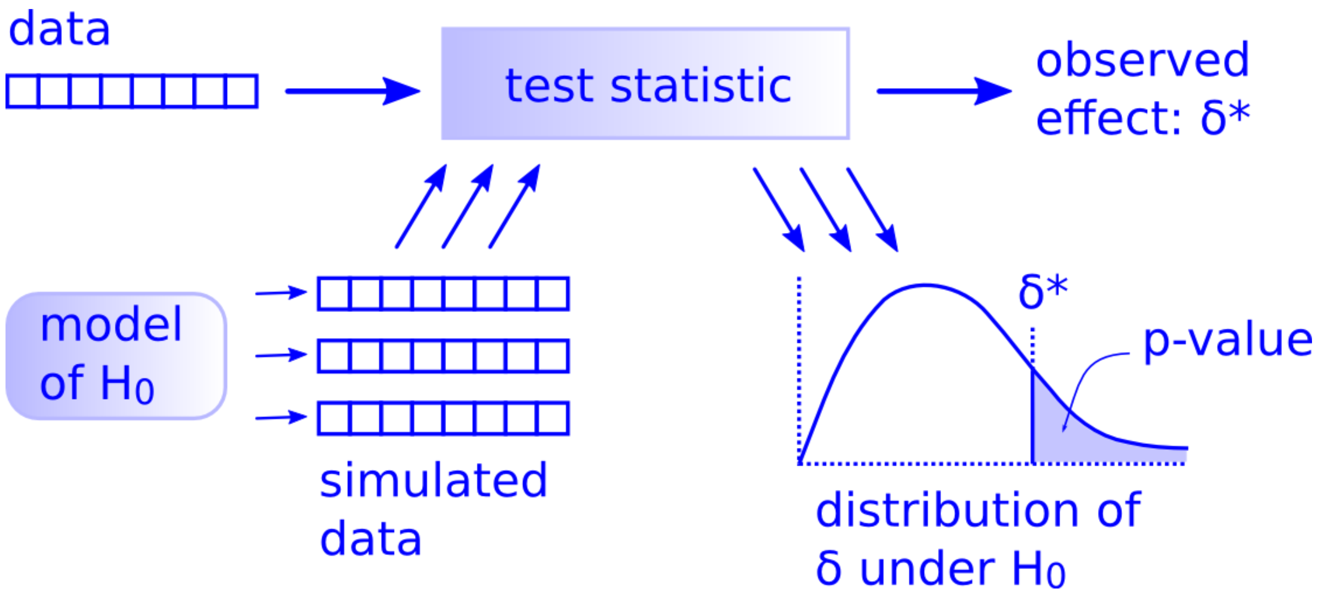
```
pnull(obs_stat, null_dist, lower.tail = FALSE)
```

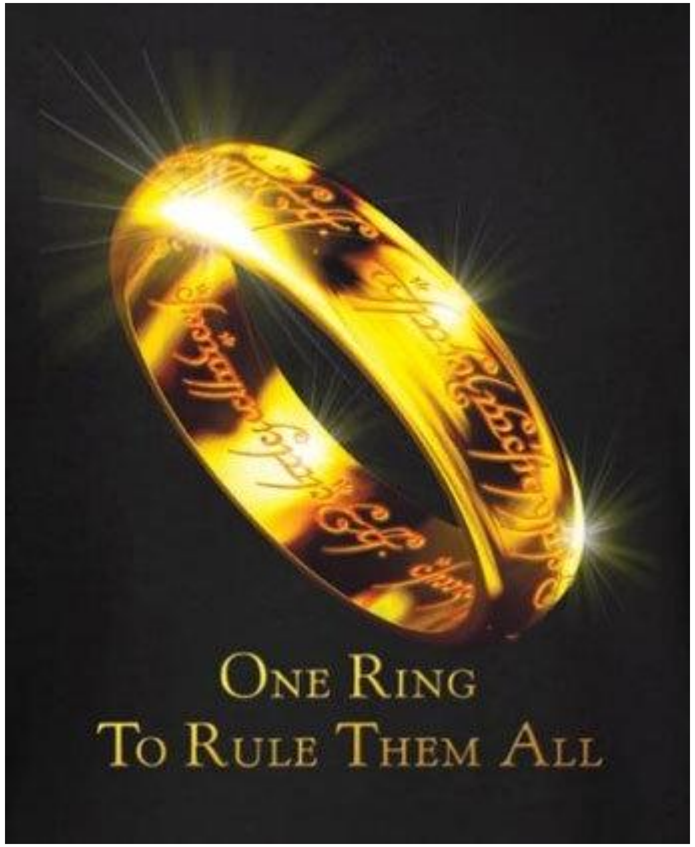
Let’s try it in R!

Hypothesis tests for correlation

The logic of hypothesis tests

There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

Hypothesis tests for correlation

Is there a positive correlation between the amount of sugar in a cereal and the number calories?

What is the population parameter and the statistic of interest?

Significance tests for correlation

Let's look at data from 30 randomly selected cereals

	Calories	Sugar
AppleJacks	117	15.0
Boo Berry	118	14.0
Cap’n Crunch	144	16.0
Corn Flakes	101	3.0

What is the first step we should do for running a hypothesis test?

Hypothesis testing for correlation

1. Write down the null and alternative in symbols and words

2. Load the data and compute the observed statistic:

```
load("cereal.Rda")
```

3. Let’s extract the calories and carbohydrates from the data frame

```
calories <- cereal$Calories
carbs <- cereal$Carbs
```

Let’s try it in R!

Step 2: What is the observed statistic?

- Also say whether you think you will be able to reject the null hypothesis based on a plot of your data

Step 3: Create the null distribution

- To start with: how we can create one point in the null distribution?
    - Hint: think about shuffling the data

Step 4: What is the p-value that you get?

Step 5: What decision would you make?

Homework 7 – Part 4: 1969 Vietnam Draft

Was the 1969 Vietnam War draft "fair" in the sense that everyone, regardless of the date they were born, was equally likely to be drafted?

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12									242	72	66	314
13									175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The first date picked was Sept 14 (sequential number 258)

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	35	135
21	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	56	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The second date picked was April 24th (sequential number 115)

What is your Draft number?

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

1969 Vietnam Draft sorted by sequential date

Date	Sequential date	Draft number
Jan 1	1	305
Jan 2	2	159
Jan 3	3	251
Jan 4	4	215
Jan 5	5	101
Jan 6	6	224
Jan 7	7	306
Jan 8	8	199
Jan 9	9	194

1969 Vietnam Draft

In a perfectly “fair”, random lottery, what should be the value of the correlation coefficient between **draft number** and **sequential date** of birthday?

Homework 7

Use hypothesis testing to assess whether there is a correlation between sequential date and draft number

- i.e., was the draft really random?

Due 11pm on Sunday November 2nd