Statistics Review

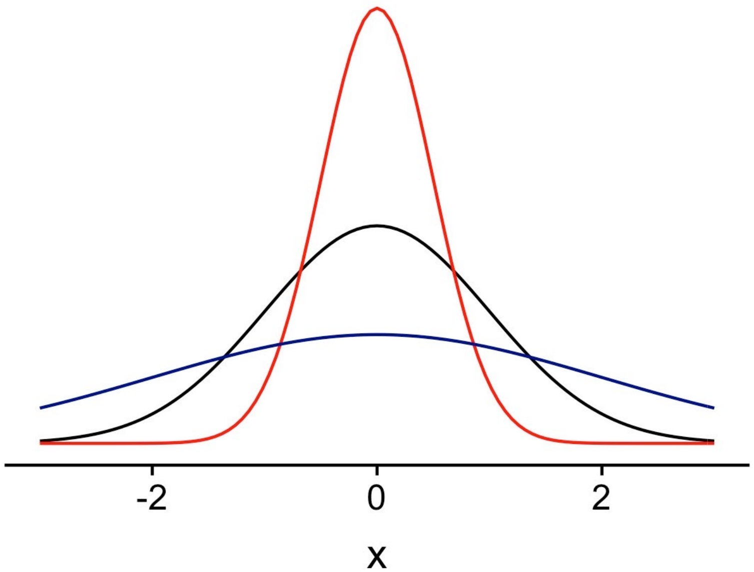
Basic Summary Values

Measures of Central Tendency

- Mean - average
- Median – central value
- Mode – most repeated value

Measures of Spread

- Range – difference between the highest and lowest value
- Standard deviation – measures the dispersion of the data



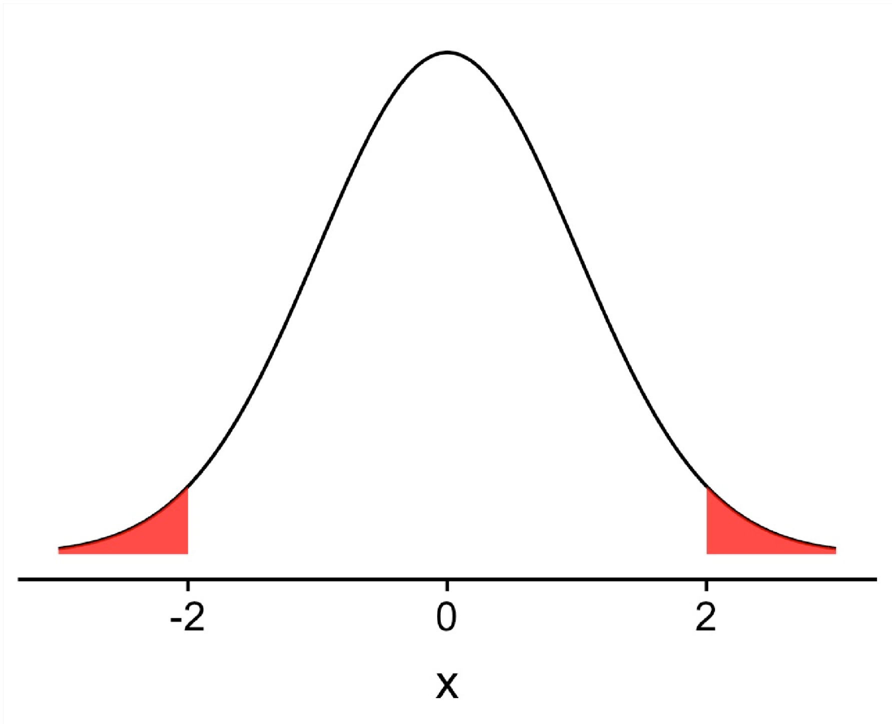
Hypothesis Testing

- Hypothesis testing compares your data to a pre-determined null distribution (usually the normal distribution). You state a null and alternative hypothesis and calculate the probability your observations happened ***under the null hypothesis.***

- Null hypothesis, **H0**: Everything happened by random chance.

- Alternative hypothesis, **H1**: My observations happened because of my idea.

- Saying p-value = 0.05 means that there's a 5% chance the observation happened randomly under the null distribution.



Test for Continuous Data: one sample t-test

- For testing continuous values against some known mean

- I have an iris with a sepal length of 7 inches and I think that it's because of my new iris fertilizer. Is that iris' sepal length abnormally large?

- **H0:** There's nothing different about the fertilizer.
- **H1:** The fertilizer does increase iris sepal length.

```
> t.test(iris$Sepal.Length, mu = 5.8) One

Sample t-test

data: iris$Sepal.Length t = 0.64092, df =
149, p-value = 0.5226 alternative
hypothesis: true mean is not equal to 5.8

95 percent confidence interval:

 5.709732 5.976934 sample

estimates:
mean of x

 5.843333
```

Test for Continuous Data: two sample t-test

- For testing 2 continuous values against each other
- Is there a difference between the sepal lengths of versicolor and virginica irises?
  - **H0:** There's no difference in the mean sepal lengths.
  - **H1:** There is a difference in the mean sepal lengths.

```
> t.test(iris[iris$Species == 'versicolor',1],
iris[iris$Species == 'virginica', 1])

Welch Two Sample t-test

data:  iris[iris$Species == "versicolor", 1]
and iris[iris$Species == "virginica", 1]

t = -5.6292, df = 94.025, p-value = 1.866e-07

alternative hypothesis: true difference in
means is not equal to 0

95 percent confidence interval:

 -0.8819731 -0.4220269

sample estimates:

mean of x mean of y

    5.936     6.588
```

Test for Continuous Data: paired two sample t-test

- For testing 2 continuous values against each other *when there is some natural pairing between the samples*
- The sleep dataset in R has data on the amount of time patients sleep on two different sleep medications compared to control. Is there a difference between the two medications?
  - **H0:** There is no difference in the amount of time patients sleep.
  - **H1:** There is a difference in the amount of time patients sleep.

```
> t.test(extra ~ group, data = sleep,
             paired = TRUE)

Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

Test for Discrete Data: chi-square

Observed Value *(aka from your sample)*

Expected Value *(aka what you already know or thought to be true before your study)*

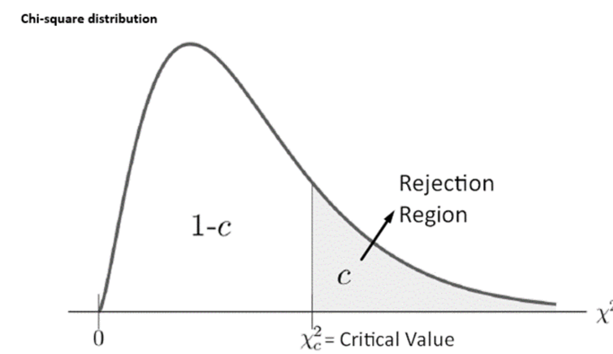
Chi-squared symbol

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

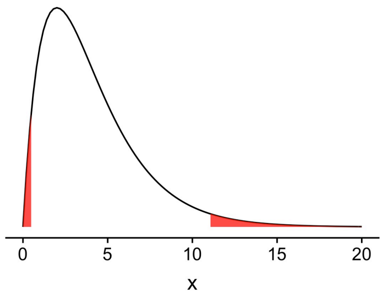
Square the result to get positive values only

Sigma *(sum/add everything up)*

Expected Value



Test for Discrete Data: chi-square



- Test for when you have counts of discrete data; test expected counts against observed counts
- Are babies more likely to be born on one day of the week over other days of the week?
  - **H0:** There is an equal chance of babies being born every day
  - **H1:** There isn't an equal chance of babies being born every day

```
> chisq.test(birth_days$num_births,
p = birth_days$exp_prob_birth)

Chi-squared test for given
probabilities

data:  birth_days$num_births
X-squared = 15.057, df = 6, p-value
= 0.01982
```

The Multiple Testing Problem

- If you do enough tests, you expect to see significant results, just *by random chance*
- Say you flip a coin 10 times and record the number of heads you get. Then you repeat the "experiment" 10 times. You expect to get about heads about 5 times

5 5 6 4 2 4 5 5 4 4

- Now let's do it 100 times

5 6 5 4 5 7 6 5 5 5 5 5 7 3 5 6 6 4 5 6

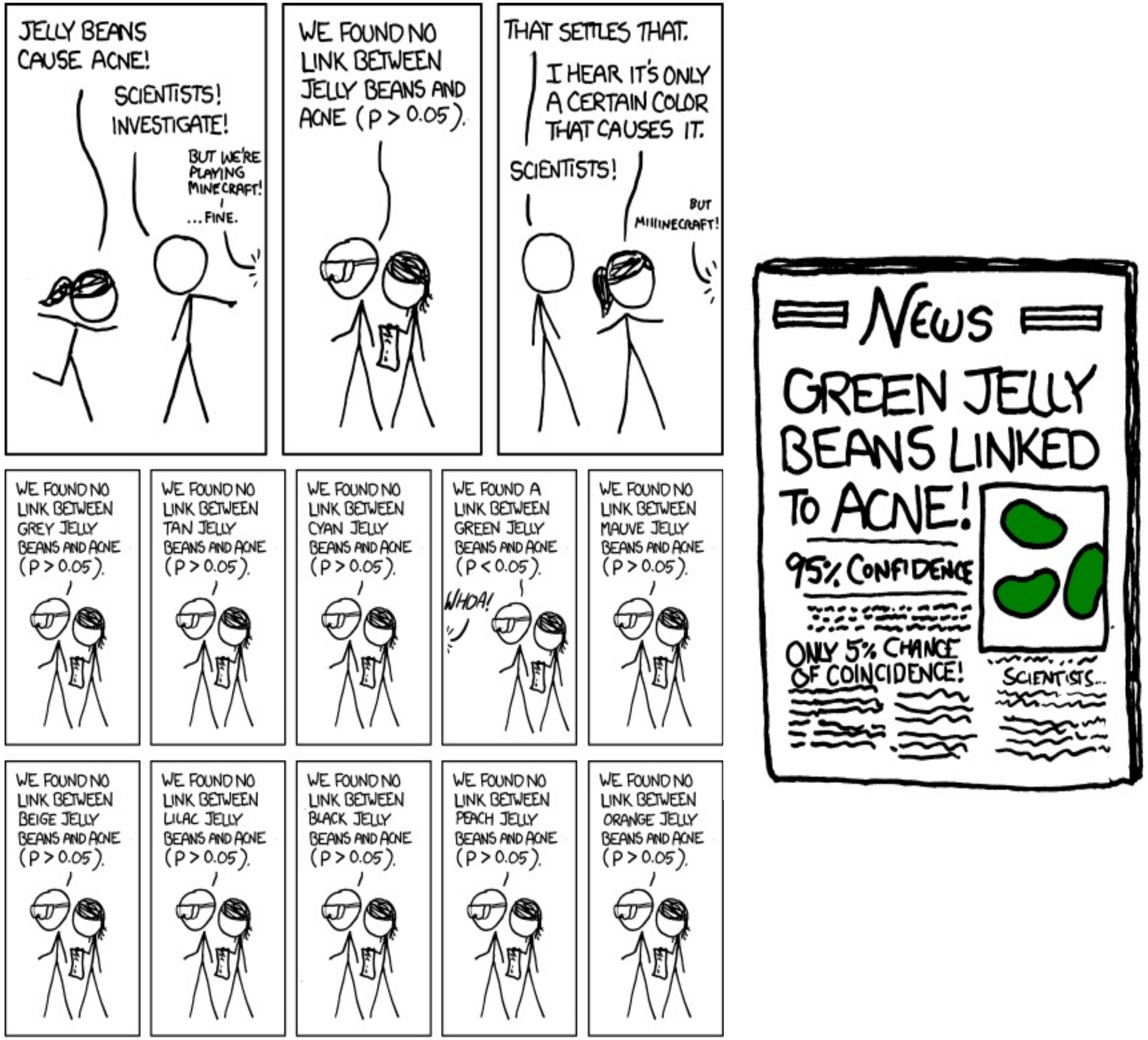
4 3 6 5 6 5 5 6 6 2 5 5 3 6 **9** 6 6 3 6 4

6 5 3 3 4 2 4 4 4 4 7 7 4 3 7 3 3 1 6 4

5 6 3 4 5 6 4 **8** 5 5 7 2 4 4 7 6 4 3 5 5

4 4 7 4 5 4 3 4 5 4 **8** 5 6 2 6 6 4 5 3 7

- Have to correct for multiple testing when you test, for example, all 20,000 genes in the human genome for differences



Pairwise Test for Multiple Conditions: ANOVA

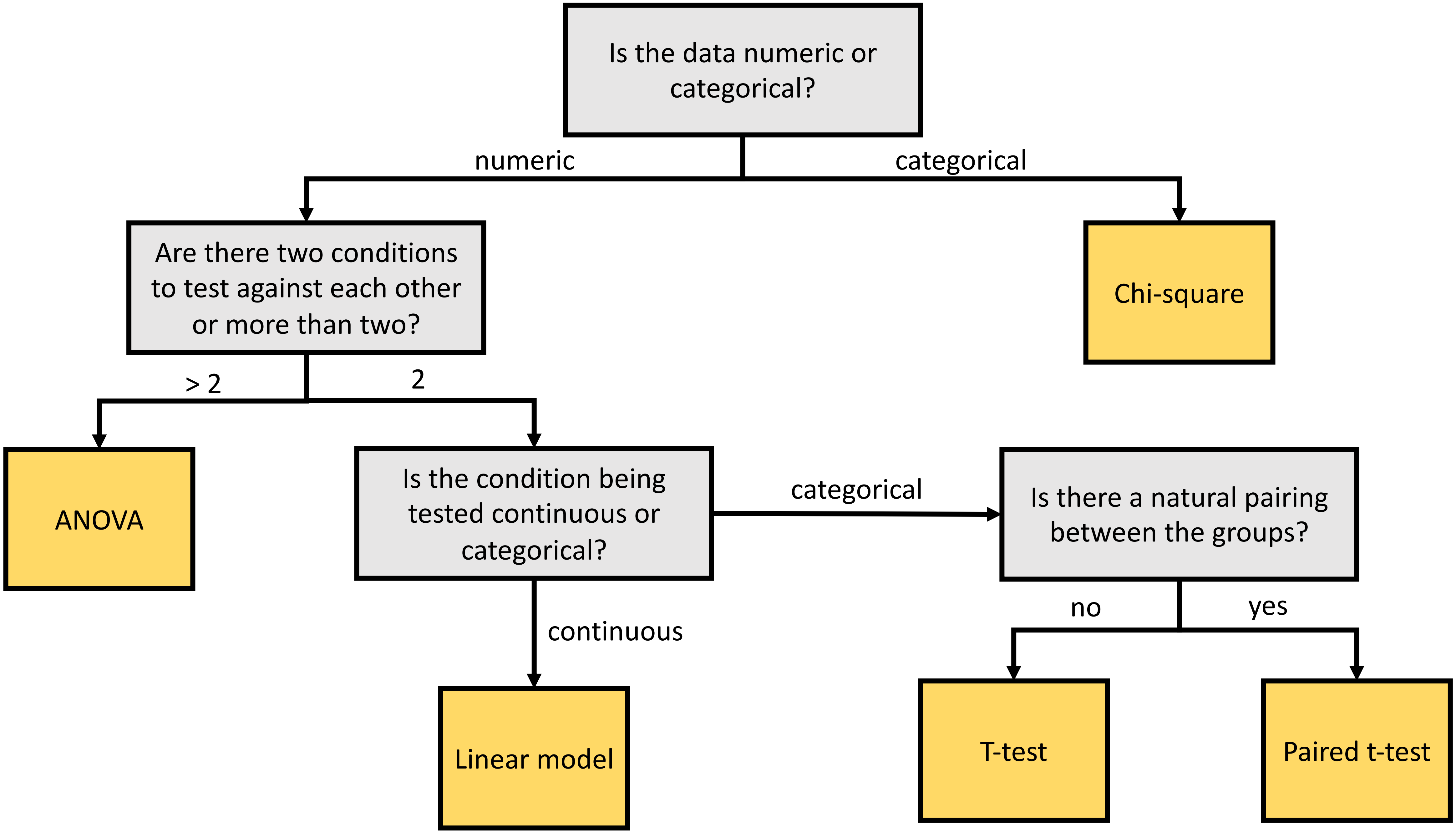
- For testing more than continuous values against all combinations of each other
- Is there a difference in sepal length between the three species of iris in the iris dataset?
  - **H0:** There is no difference
  - **H1:** There is a difference between at least one group

```
> aov(Sepal.Length ~ Species, data = iris) %>%
TukeyHSD()

  Tukey multiple comparisons of means

    95% family-wise confidence level


Fit: aov(formula = Sepal.Length ~ Species, data =
iris)


$Species

                      diff       lwr       upr p adj

versicolor-setosa    0.930 0.6862273 1.1737727     0

virginica-setosa     1.582 1.3382273 1.8257727     0

virginica-versicolor 0.652 0.4082273 0.8957727     0
```

Test for Continuous Conditions: Linear Model

- For testing continuous variables over a continuous condition (like DNA methylation over time)
- AKA finding a line of best fit
- Is there an association between sepal width and sepal length in the iris dataset?
  - **H0:** There is no relationship
  - **H1:** There is a relationship

```
> lm(Sepal.Length ~ Sepal.Width, data = iris)
Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)


Coefficients:
(Intercept)  Sepal.Width
     6.5262      -0.2234
```


Is the data numeric or categorical?
numeric
categorical
Are there two conditions to test against each other or more than two?
Chi-square
> 2
2
ANOVA
Is the condition being tested continuous or categorical?
categorical
Is there a natural pairing between the groups?
continuous
no
yes
Linear model
T-test
Paired t-test

Demo

Statistical interference using Resampling

- jackknife

- bootstrap

- permutation

- cross validation

Permutation

Group comparison
- get p-value of statistic

Group A		Group B	
27	24	21	13
20	29	22	22
21	18	15	20
26	20	12	24
27	17	21	18
31	31	16	20
24	20	19	23
21	25	15	19
20	28	22	24
19	21		
23	27		
24	28		
19			

Permutation

Group comparison
 - get p-value of statistic

Group A

```
> A=c(27,24,20,29,21,18,26,20,27,17,31,31,24,20,21,25,20,28,19,21,23,27,24,28,19)
> mean(A)
[1] 23.6
```

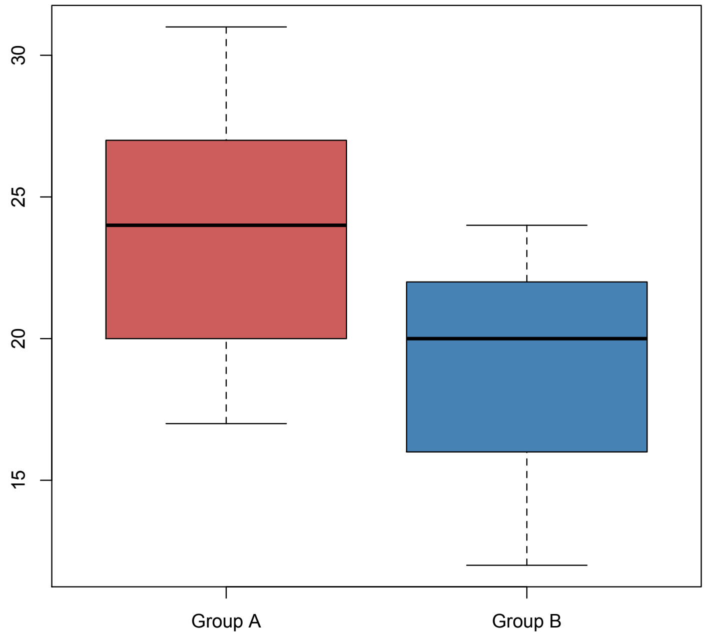
Group B

```
> B=c(21,13,22,22,15,20,12,24,21,18,16,20,19,23,15,19,22,24)
> mean(B)
[1] 19.2
```

Mean difference A vs B: + 4.4

Are Group A and B different?

Is difference 4.4 statistically different?



There are 2 possible way to determine:

Are Group A and B different?

There are 2 possible way to determine:

1. Classical statistics (analytical method)
2. Computational method

1. Classical statistics (analytical method)
STAT 101

Student's *t*-test

Equal or unequal sample sizes, unequal variances

Welch's t-*test*

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

1. Classical statistics (analytical method)
STAT 101

Student's *t*-test

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

1. Classical statistics (analytical method)
STAT 101

Student's *t*-test

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

1. Classical statistics (analytical method)
STAT 101

Student's *t*-test

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

$$\text{d.f.} = \frac{(4.1^2/25 + 3.7^2/18)^2}{(4.1^2/25)^2/(25-1) + (3.7^2/18)^2/(18-1)}$$

$$\text{d.f.} = 39.1$$

1. Classical statistics (analytical method)
STAT 101

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

$$\text{d.f.} = \frac{(4.1^2/25 + 3.7^2/18)^2}{(4.1^2/25)^2/(25-1) + (3.7^2/18)^2/(18-1)}$$

$$\text{d.f.} = 39.1$$

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
...											
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

1. Classical statistics (analytical method)
STAT 101

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

$$\text{d.f.} = \frac{(4.1^2/25 + 3.7^2/18)^2}{(4.1^2/25)^2/(25-1) + (3.7^2/18)^2/(18-1)}$$

$$\text{d.f.} = 39.1$$

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
...											
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.68[illegible]	2.021	[illegible].423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

3.67 > 2.021

1. Classical statistics (analytical method)
STAT 101

$$t = \frac{23.6 - 19.2}{\sqrt{\frac{17.1}{25} + \frac{13.5}{18}}} = 3.67$$

$$\text{d.f.} = \frac{(4.1^2/25 + 3.7^2/18)^2}{(4.1^2/25)^2/(25-1) + (3.7^2/18)^2/(18-1)}$$

$$\text{d.f.} = 39.1$$

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
...											
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496

3.67 > 2.021 ✓

1. Classical statistics (analytical method)
STAT 101

Difference 4.4 is statistically significant **p<0.05** level

```
> t.test(A,B)

	Welch Two Sample t-test

data:  A and B
t = 3.6582, df = 39.113, p-value = 0.0007474
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.957472 6.798084
sample estimates:
mean of x mean of y 
 23.60000  19.22222 
```

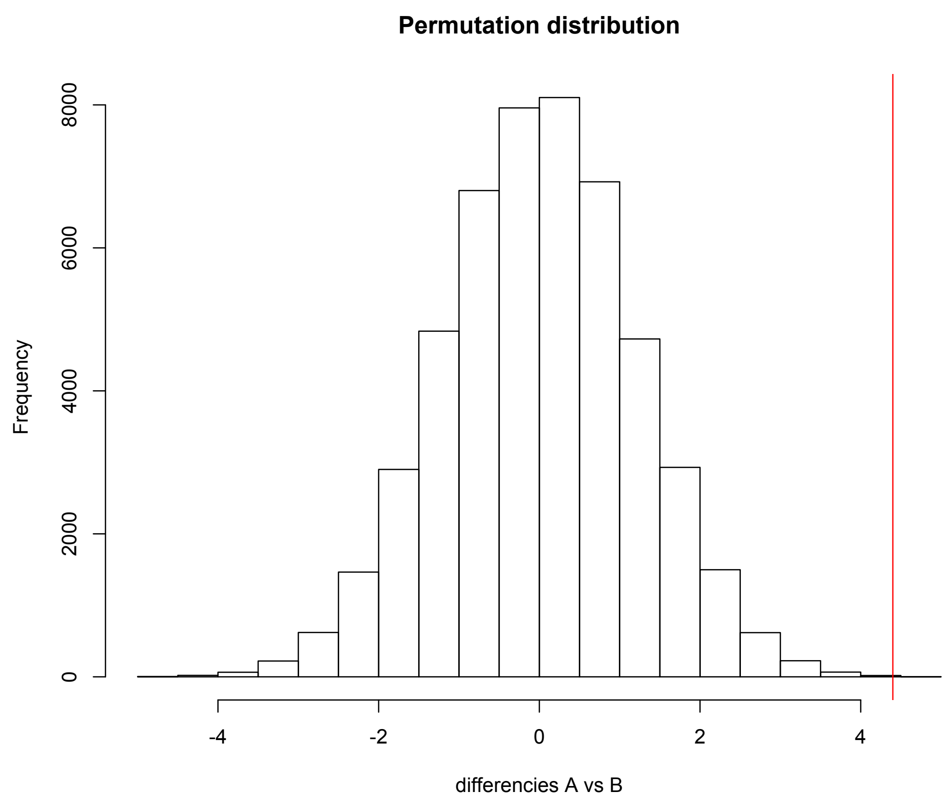
2. Computational method

Group A		Group B	
27	24	21	13
20	29	22	22
21	18	15	20
26	20	12	24
27	17	21	18
31	31	16	20
24	20	19	23
21	25	15	19
20	28	22	24
19	21		
23	27		
24	28		
19			

2. Computational method

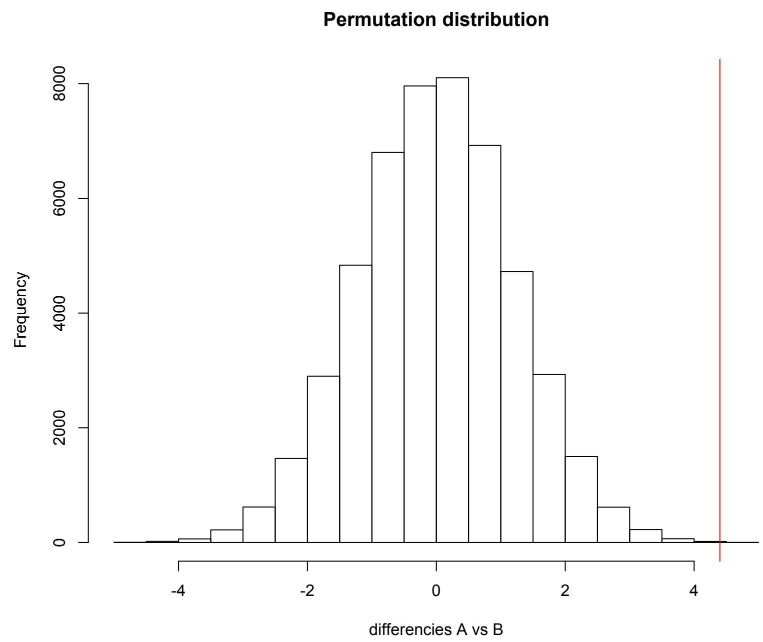
Group A		Group B	
22	24	28	13
20	29	22	27
21	16	15	20
26	20	27	24
12	17	21	24
31	31	18	20
18	23	19	20
21	25	15	19
20	28	19	24
22	21		
23	27		
24	21		
19			

differencies in mean 2.7


Permutation distribution
Frequency
0
2000
4000
6000
8000
-4
-2
0
2
4
differencies A vs B

2. Computational method

1. Ability to follow logical statement
2. Random number generator - in R sample() command
3. Iteration



p-val: **0.00075**

7-8 of of 10,000

~**75** out of 100,000