浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

The mathematics of the *t*-test

Why and when does it work?

ADS 2, Lecture 9

Rob Young – robert.young@ed.ac.uk

Semester 1, 2019/20

This statistics lecture contains trade secrets!

- William Sealy Gosset, Head Experimental Brewer at Guinness
- Developed statistical techniques to assess quality of the finished product based on sampling during production.
- He published his 1908 paper under the pseudonym ‘Student’ which is where the Student in Student’s *t*-distribution comes from.

Rationale for the *t*-test

- What do we mean when we say two values are ‘different’?

- Is the difference in values greater than what we might expect?

Rationale for the *t*-test

- What do we mean when we say two values are ‘different’?
- Is the difference in values greater than what we might expect?
- Compare the sample mean to a known value or distribution.
  - Null hypothesis: the values are equal, they come from the same population
- Appropriate for small sample sizes ($n < 100$).
- Generally, we assume a normal distribution of the underlying population.

Learning objectives

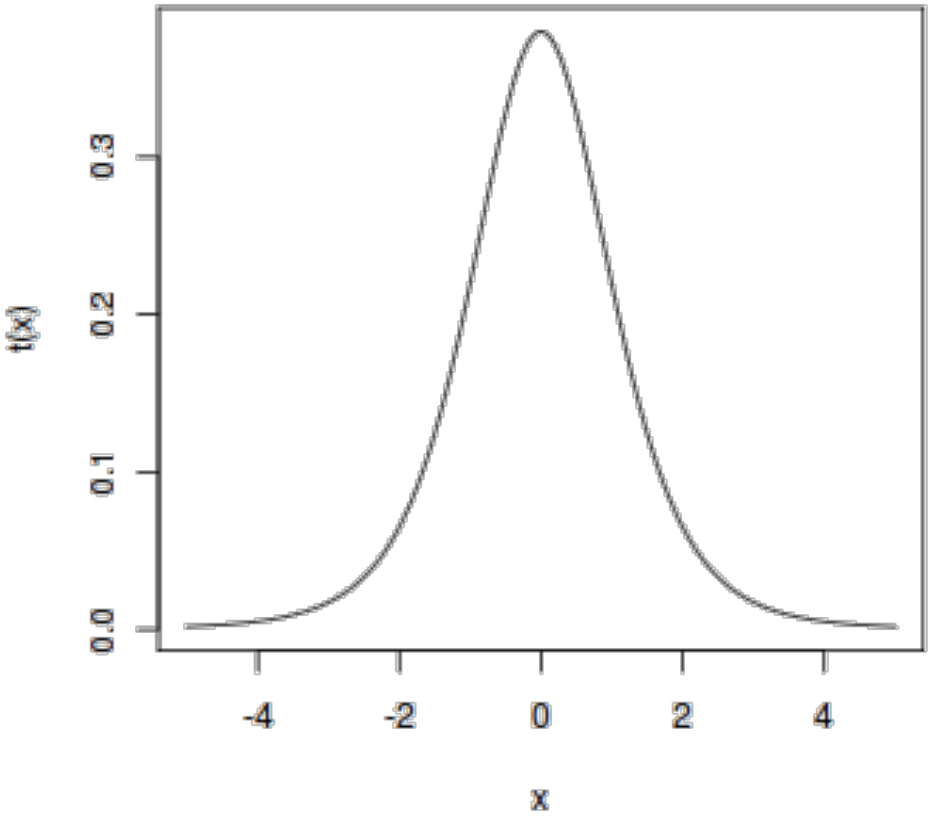
After this lecture, you should be able to:

- **Understand the mathematics behind the *t*-test.**

- Use the Student's *t*-distribution to determine the significance of a given sample.

- Describe the assumptions that need to be met to apply the *t*-test appropriately.

Student's *t*-distribution



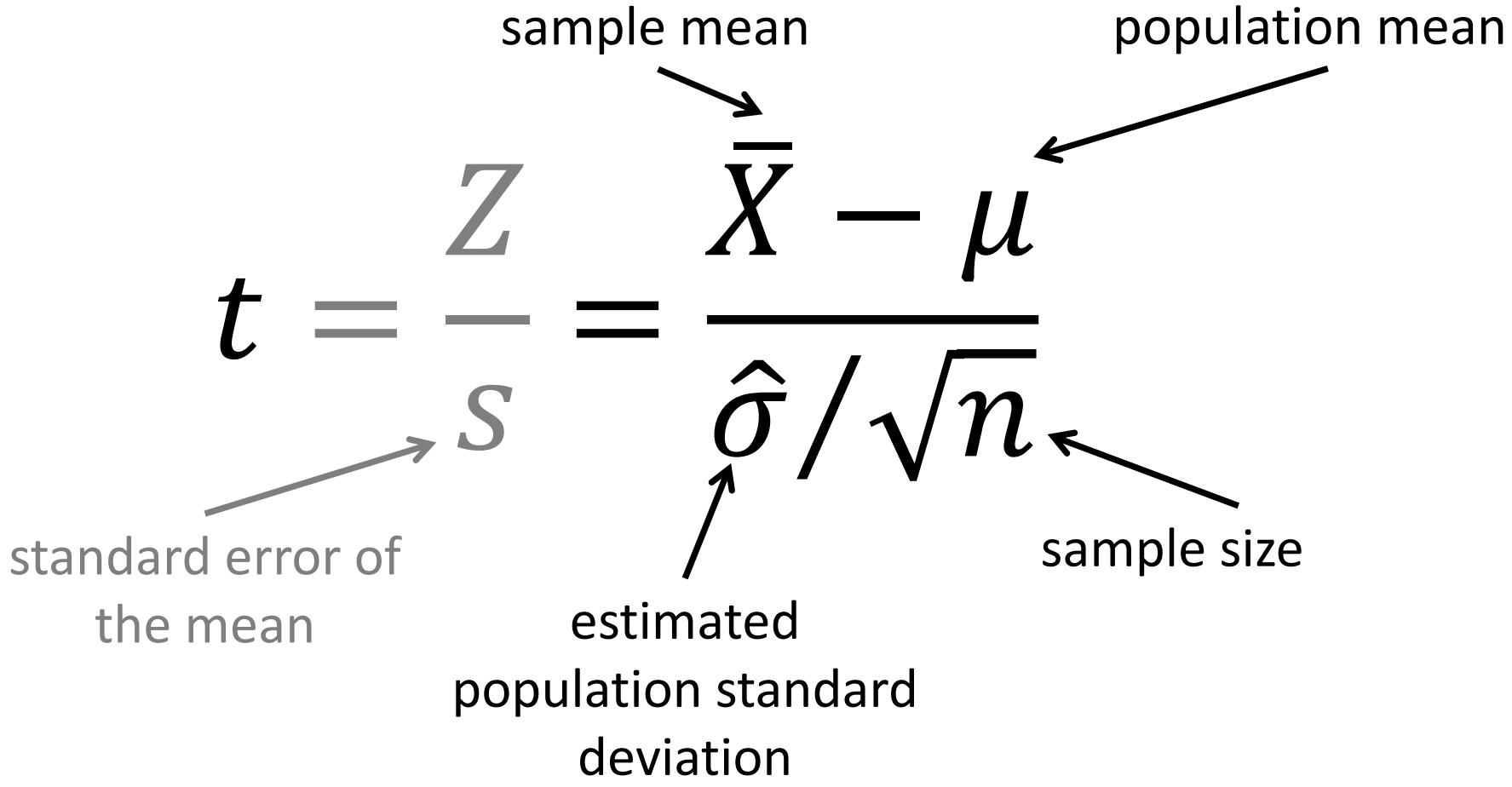
- Continuous probability distribution.
- Symmetric and bell-shaped.
- Derived from a small sample size where the population standard deviation is unknown.

The only equation in this lecture: How to calculate the *t*-statistic

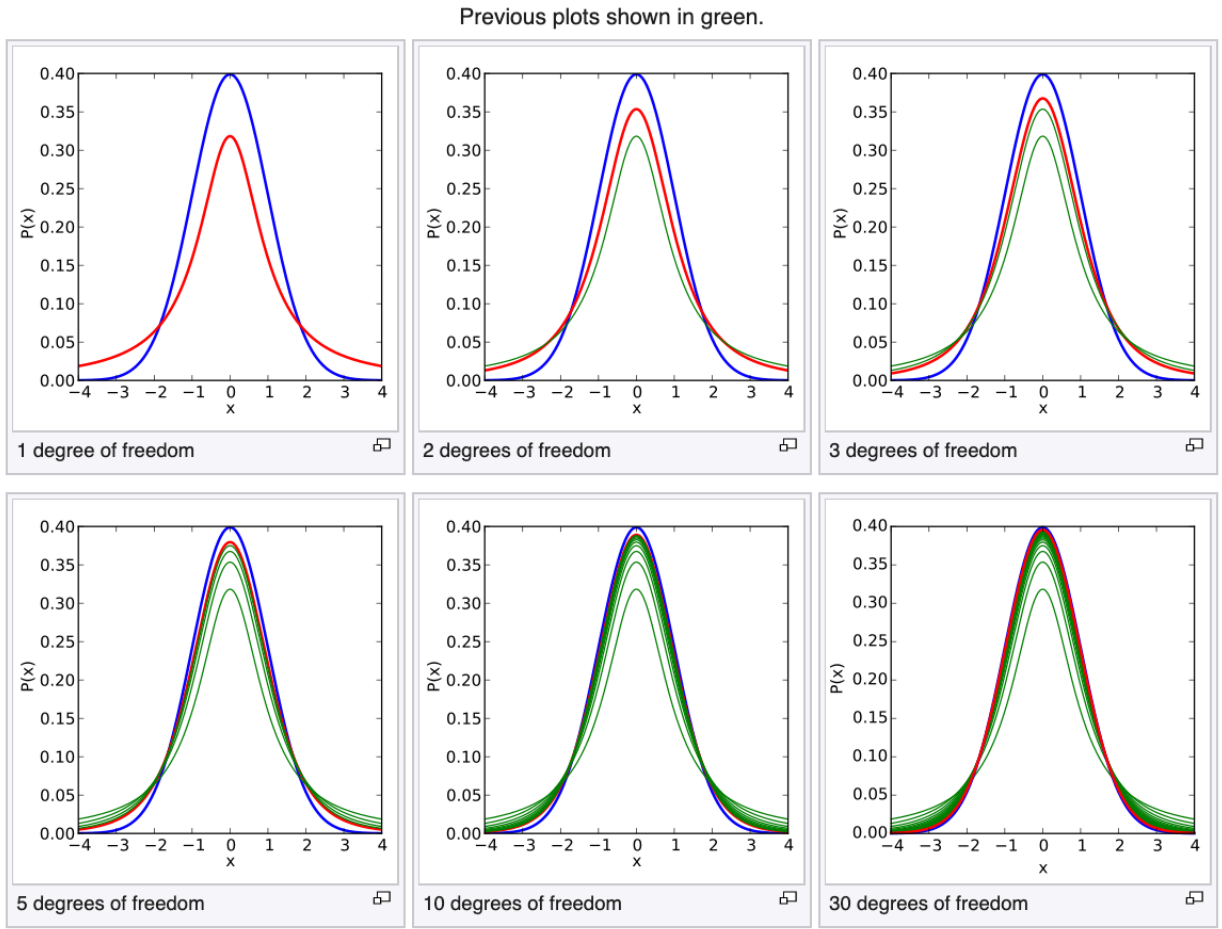
$$t = \frac{Z}{s} = \frac{\text{Difference in values}}{\text{Uncertainty in values}}$$

standard error of the mean

The only equation in this lecture: How to calculate the *t*-statistic



Student's *t*-distribution: approaching the normal distribution



- **Symmetric and bell-shaped**.
- Heavier tailed compared to the **normal distribution**.
- Distribution depends on one degree of freedom (the mean) rather than the mean and standard deviation in the normal distribution (df = 2).

So just what are these degrees of freedom?

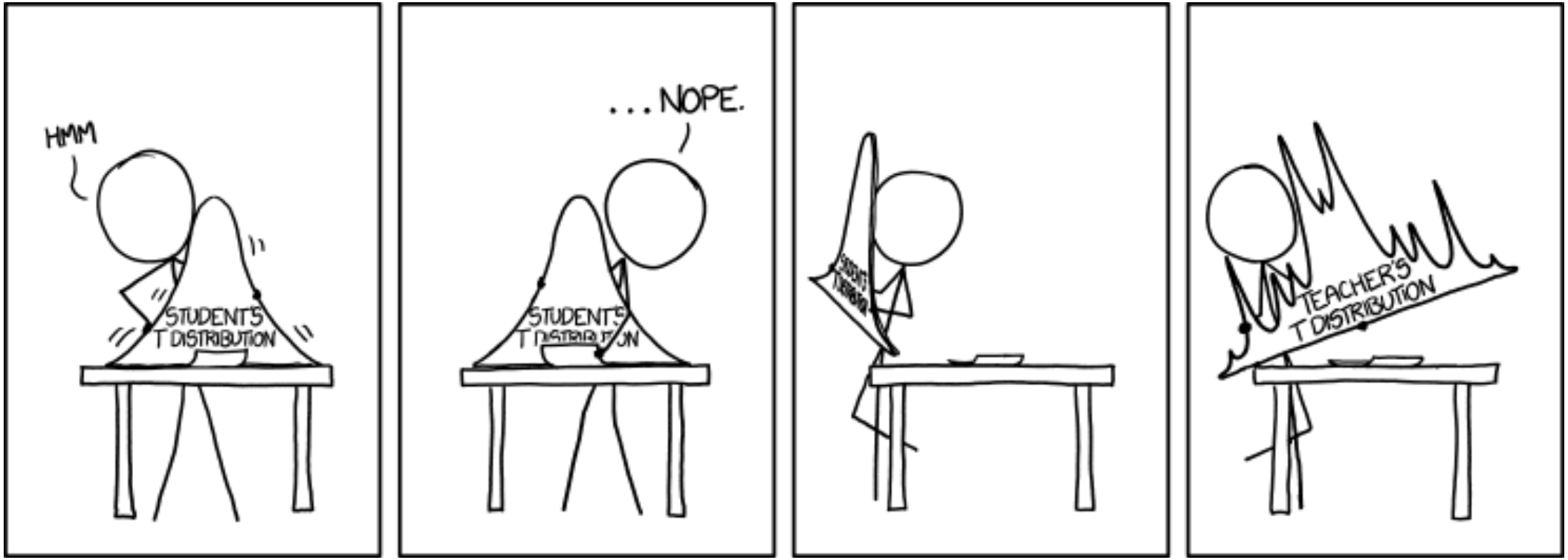
"The degrees of freedom in a statistical calculation represent how many values involved in a calculation have the freedom to vary."

Usually (always, in my experience!), this means…

$$n - 1$$

sample size

Let’s have a break for 2 minutes to enjoy this cartoon



Learning objectives

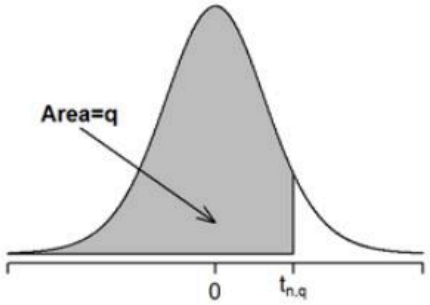
After this lecture, you should be able to:

- Understand the mathematics behind the *t*-test.

- **Use the Student's *t*-distribution to determine the significance of a given sample.**

- Describe the assumptions that need to be met to apply the *t*-test appropriately.

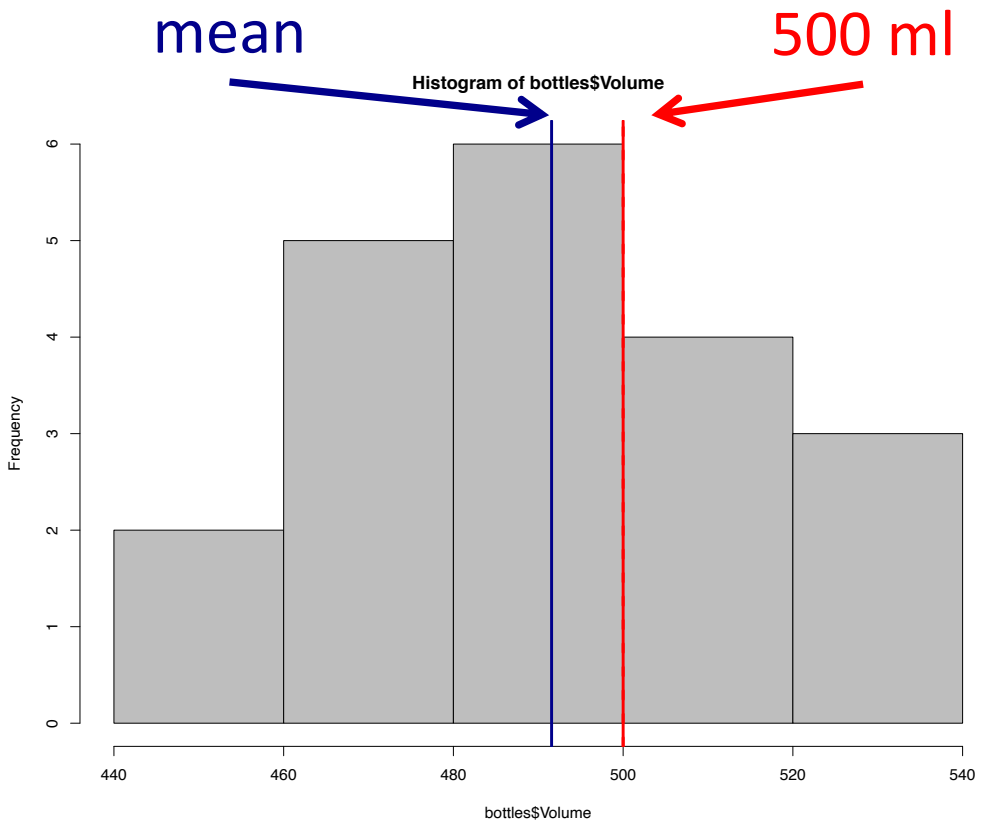
Critical values of the Student's *t*-distribution



	q = 0.6	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
n = 1	0.3249	1.0000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.2887	0.8165	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.2767	0.7649	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.2707	0.7407	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.2672	0.7267	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.2648	0.7176	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.2632	0.7111	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.2619	0.7064	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.2610	0.7027	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.2602	0.6998	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.2596	0.6974	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.2590	0.6955	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.2586	0.6938	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.2582	0.6924	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140

- Each *t*-value has an associated probability, determined by the degrees of freedom.
- Critical values can be looked up in Statistical Tables like this.

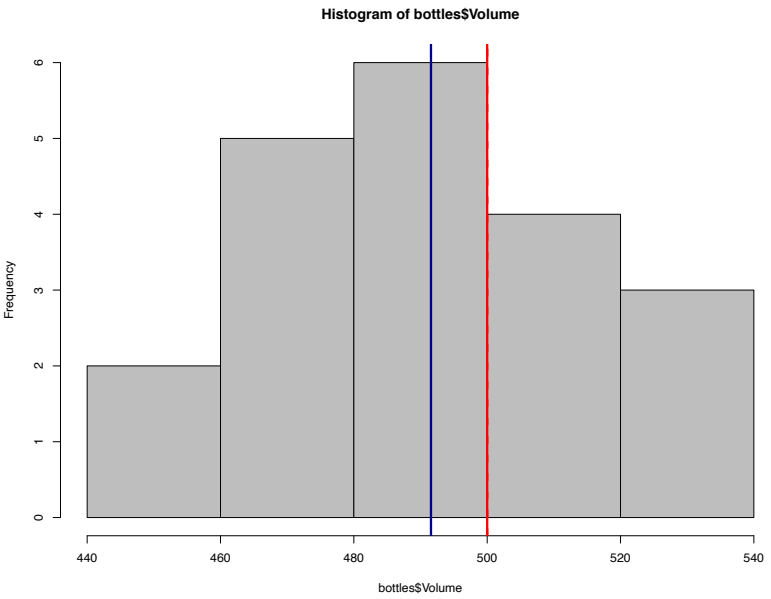
Is the factory filling each bottle with enough Guinness?



- 10 million glasses are drunk (and assumed to be brewed) every day, so we can't check them all!
- The volume of 20 bottles have been measured.
- The mean volume is 491.6 ml, but the required volume is 500 ml.

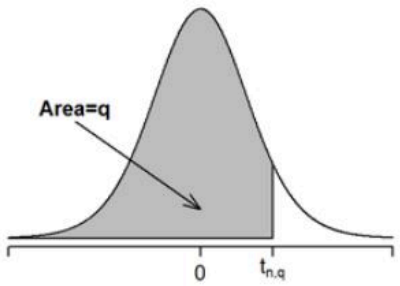
The only equation in this lecture: Calculating the *t*-statistic

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{491.6 - 500}{24.8/\sqrt{20}} = -1.52$$



(I calculated this value in advance of the lecture)

Is this *t*-statistic significant?

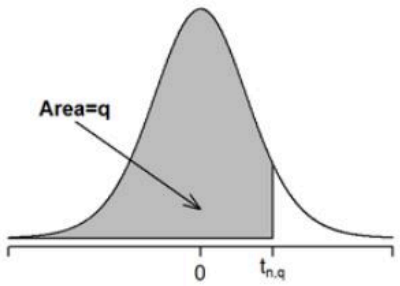


d.f.	$t_{.100}$	$t_{.050}$*	$t_{.025}$**	$t_{.010}$	$t_{.005}$	d.f.
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

There is only a 5% probability that a sample with 10 degrees of freedom will have a t value greater than 1.812.
* one tail 5% α risk ** two tail 5% α risk

- Observed *t* = -1.52.
- How many degrees of freedom (20 bottles)?

Is this *t*-statistic significant?

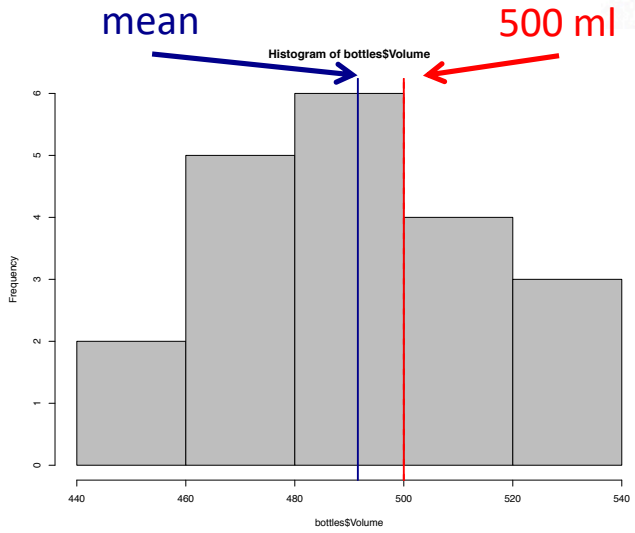


d.f.	$t_{.100}$	$t_{.050}$*	$t_{.025}$**	$t_{.010}$	$t_{.005}$	d.f.
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

There is only a 5% probability that a sample with 10 degrees of freedom will have a t value greater than 1.812.
* one tail 5% α risk ** two tail 5% α risk

- Observed *t* = -1.52.
- 19 degrees of freedom (n = 20).

Is this *t*-statistic significant?

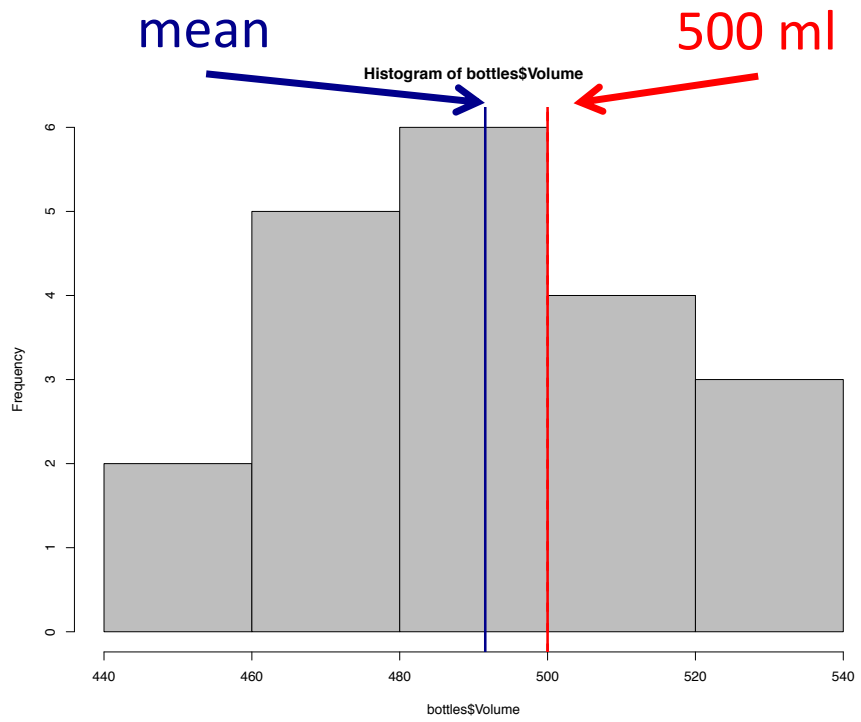


d.f.	$t_{.100}$	$t_{.050}$*	$t_{.025}$**	$t_{.010}$	$t_{.005}$	d.f.
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

There is only a 5% probability that a sample with 10 degrees of freedom will have a t value greater than 1.812.
* one tail 5% α risk ** two tail 5% α risk

- Observed t = -1.52.
- 19 degrees of freedom (n = 20).
- Assume a threshold one-tailed p-value 0.05 → critical t = 1.729.
- Is the factory filling each bottle with enough Guinness?

In R, use t.test(SAMPLE, mu = VALUE)




```
> t.test(bottles$Volume, mu = 500)

        One Sample t-test

data:  bottles$Volume
t = -1.5205, df = 19, p-value = 0.1449
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 479.9667 503.1743
sample estimates:
mean of x
 491.5705

>
```

- Much easier than looking up tables of critical values!
- Reports a *p*-value, here it is $p = 0.14$.

Learning objectives

After this lecture, you should be able to:

- Understand the mathematics behind the *t*-test.

- Use the Student's *t*-distribution to determine the significance of a given sample.

- **Describe the assumptions that need to be met to apply the *t*-test appropriately.**

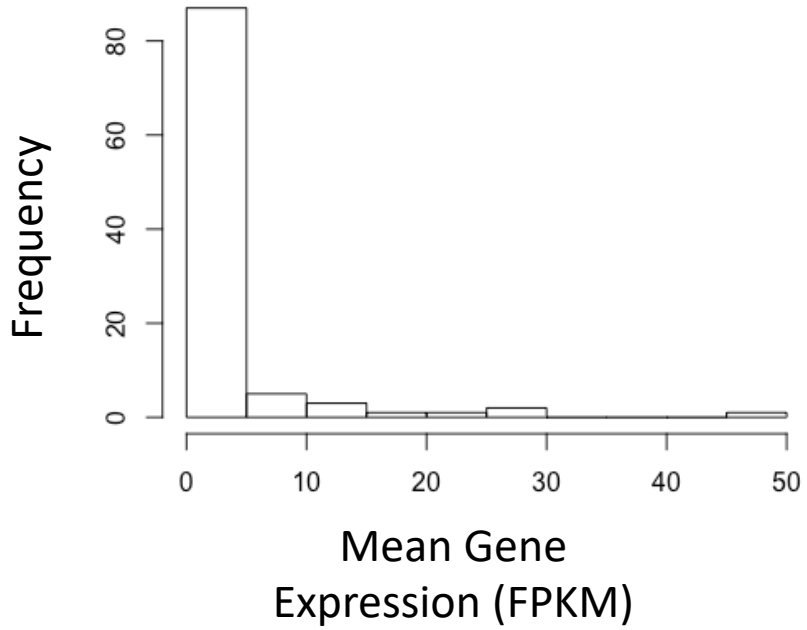
Assumptions required for using the *t*-statistic

1. Data is continuous and randomly-selected.
   → see lecture 2 on sampling

2. The sampling distribution is normally distributed.
   → see lecture 3 on the central limit theorem.

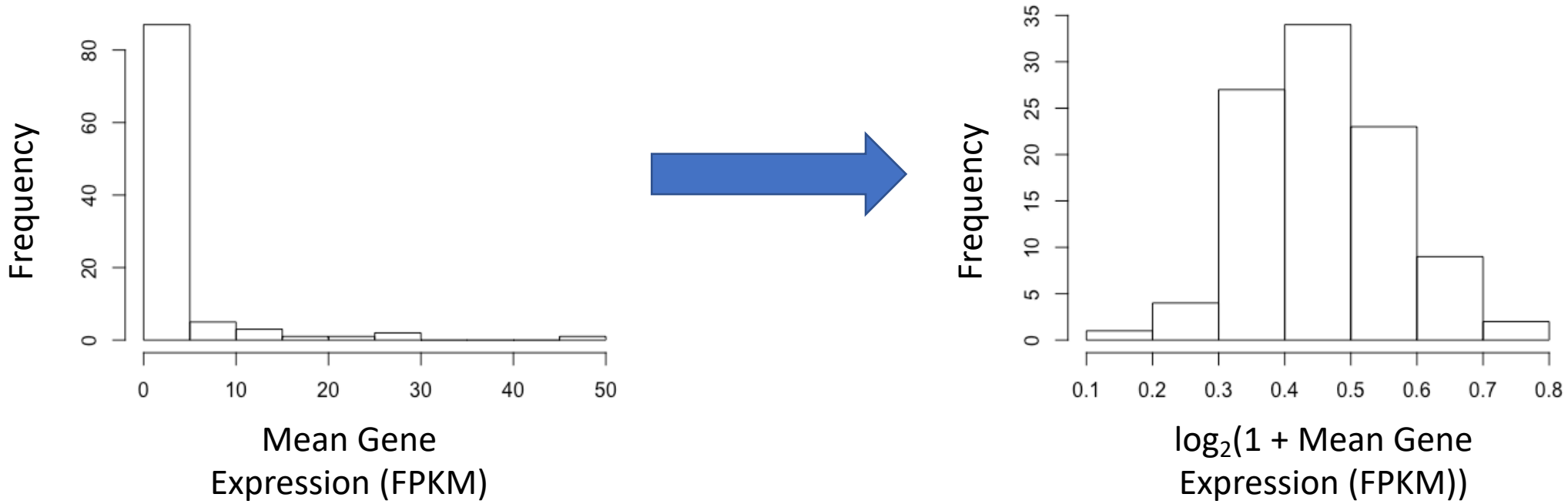
3. The mean and standard error are independent.
   → nearly always true, but could test by simulation.

Assumption 2/3: but what if my sampling distribution is not normally distributed?



- Gene expression measurements are frequently very far from the normal or *t*-distribution.

Assumption 2/3: but my sampling distribution is not normally distributed?



- Gene expression measurements are frequently very far from the normal or *t*-distribution.
- Transformations, such as the logarithmic, can make your data more ‘normal’.

Assumptions required for using the *t*-statistic

1. Data is continuous and randomly-selected.
   → see lecture 2 on sampling

2. The sampling distribution is normally distributed.
   → see lecture 3 on the central limit theorem.

3. The mean and standard error are independent.
   → nearly always true, but can be tested (problem sheet).

Learning objectives

Now you should be able to:

- Understand the mathematics behind the *t*-test.

- Use the Student's *t*-distribution to determine the significance of a given sample.

- Describe the assumptions that need to be met to apply the *t*-test appropriately.

浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

The mathematics of the *t*-test

Any questions?

ADS 2, Lecture 9

Rob Young – robert.young@ed.ac.uk

Semester 1, 2019/20