Contingency analysis: associations between categorical variables

Contingency analysis

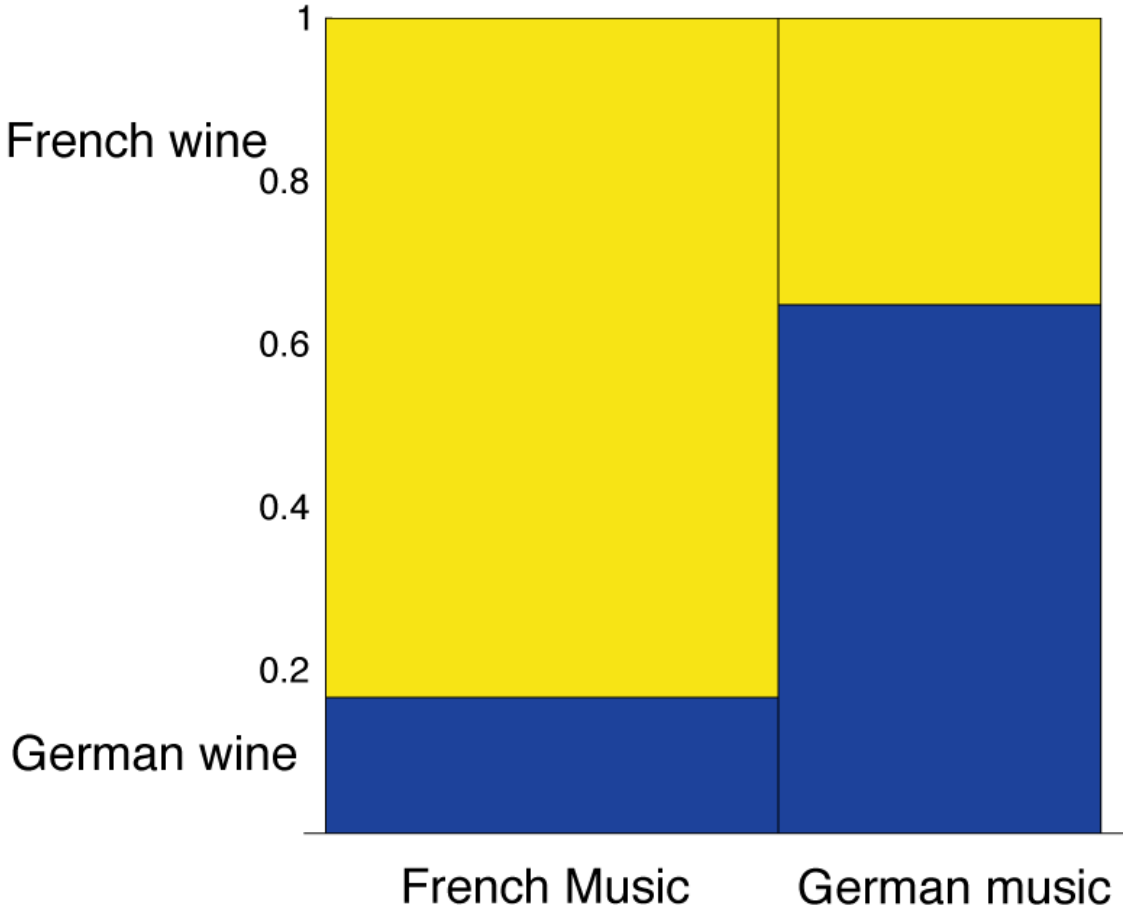
- Test the independence of two or more categorical variables

- We'll learn one kind: χ^2 contingency analysis

Music and wine buying

OBSERVED	French music playing	German music playing	Totals
Bottles of French wine sold	40	12	52
Bottles of German wine sold	8	22	30
Totals	48	34	82

Mosaic plot


1
French wine
0.8
0.6
0.4
0.2
German wine
French Music
German music

Hypotheses

- H_0: The nationality of the bottle of wine *is independent* of the nationality of the music played when it is sold.

- H_A: The nationality of the bottle of wine sold *depends* on the nationality of the music being played when it is sold.

Calculating the expectations

With independence,

Pr[French wine AND French music] =

Pr[French wine] × Pr[French music]

Calculating the expectations

EXP.	French music	German music	Totals
French wine sold			52
German wine sold			30
Totals	48	34	82

Pr[French wine] =

Pr[French music] =

If H_0 is true,
Pr[French wine AND French music] =

Calculating the expectations

EXP.	French music	German music	Totals
French wine sold			52
German wine sold			30
Totals	48	34	82

Pr[French wine] = 52/82=0.634

Pr[French music] = 48/82= 0.585

If H_0 is true,

Pr[French wine AND French music] = (0.634)(0.585) = 0.37112

Calculating the expectations

EXP.	French music	German music	Totals
French wine sold	0.37 x 82 = 30.4	21.6	52
German wine sold	17.6	12.4	30
Totals	48	34	82

By H_0,
Pr[French wine AND French music] = (0.634)(0.585)=0.37112

χ^2

$$\chi^2 = \sum_i \frac{\left(Observed_i - Expected_i\right)^2}{Expected_i}$$

$$= \frac{(40-30.4)^2}{30.4} + \frac{(12-21.6)^2}{21.6} + \frac{(8-17.6)^2}{17.6} + \frac{(22-12.4)^2}{12.4}$$

$$= 20.0$$

Degrees of freedom

df= (# columns -1)(#rows -1)

For music/wine example,

df = (2-1)(2-1) = 1

Table A - χ^2 distribution

	α									
df	0.999	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.0000016	0.000039	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12
9	1.15	1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59	27.88
10	1.48	2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19	29.59
11	1.83	2.60	3.05	3.82	4.57	19.68	21.92	24.72	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	32.91

Conclusion

$\chi^2 = 20.0 >> \chi^2_{1,\alpha=0.05} = 3.84,$

So we can reject the null hypothesis of independence, and say that the nationality of the wine sold did depend on what music was played.

Moreover, $\chi^2 = 20.0 >> \chi^2_{1,\alpha=0.001} = 10.83$, so we can say $P < 0.001$.

Assumptions

- This χ^2 test is just a special case of the χ^2 goodness-of-fit test, so the same rules apply.

- You can't have any expectation less than 1, and no more than 20% < 5.

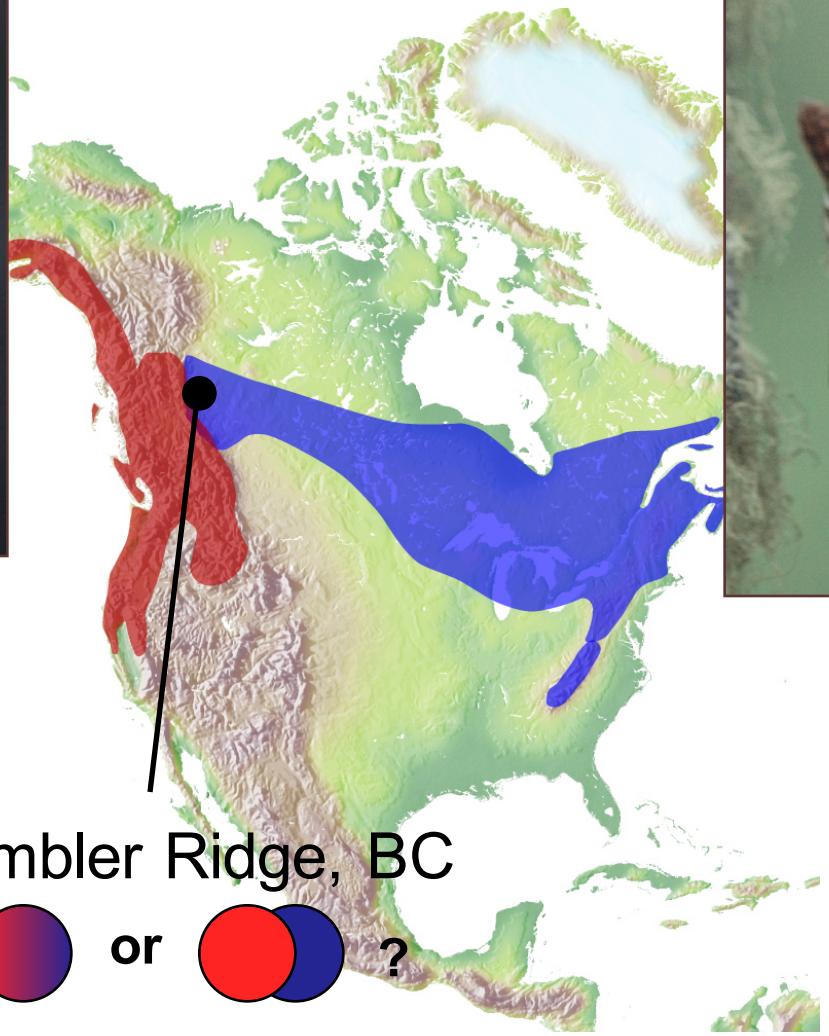
Fisher's exact test

- For 2 x 2 contingency analysis
- Does not make assumptions about the size of expectations
- JMP (or other programs) will do it, but cumbersome to do by hand

Winter Wren (*Troglodytes troglodytes*)

- Are western and eastern forms (currently considered subspecies) actually reproductively isolated, and therefore separate species?

T. (t.) pacificus



T. t. hiemalis

Photos by D. Irwin

Association of DNA and song: The winter wren contact zone

OBSERVED	Western song	Eastern song	Totals
Western mtDNA	12	0	12
Eastern mtDNA	0	4	4
Totals	12	4	16

Data from Toews & Irwin 2008, *Molecular Ecology*

Calculating the expectations

EXP.	Western song	Eastern song	Totals
Western mtDNA			12
Eastern mtDNA			4
Totals	12	4	16

A shortcut for calculating expectations (assuming H_0 is true):

$$Exp[row\ i, column\ j] = \frac{(row\ i\ total)(column\ j\ total)}{grand\ total}$$

Exp[w mtDNA, w song] = 12*12/16 = 9

Comparing observed and expected

OBS.	Western song	Eastern song	Totals
Western mtDNA	12	0	12
Eastern mtDNA	0	4	4
Totals	12	4	16

EXP.	Western song	Eastern song	Totals
Western mtDNA	9	3	12
Eastern mtDNA	3	1	4
Totals	12	4	16

Too many of the expected are below 5, so we cannot use the χ^2 contingency test. Instead, we use a computer to do Fisher's exact test:

$P = 0.00055$, so we reject the H_0 of no association.

χ^2

$$\chi^2 = \sum_i \frac{\left(Observed_i - Expected_i\right)^2}{Expected_i}$$

$$= \frac{\left(20 - 12.23\right)^2}{12.23} + \frac{\left(69 - 76.77\right)^2}{76.77} + \frac{\left(16 - 23.77\right)^2}{23.77} + \frac{\left(157 - 149.23\right)^2}{149.23}$$

$$= 8.67$$

In-class Exercise

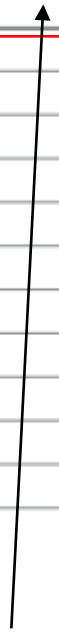
Do mosquitos infected with malaria bite more people?

$\chi^2 = 8.67$

df = (2-1)(2-1)= 1

Table A - χ^2 distribution

	α									
df	**0.999**	**0.995**	**0.99**	**0.975**	**0.95**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	0.0000016	0.000039	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.002	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60	13.82
3	0.02	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.09	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.21	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.38	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	0.60	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	0.86	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12
9	1.15	1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59	27.88
10	1.48	2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19	29.59
11	1.83	2.60	3.05	3.82	4.57	19.68	21.92	24.72	26.76	31.26
12	2.21	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30	32.91



$\chi^2 = 8.67$

One vs two categorical variables

χ^2 goodness-of-fit

one categorical variable. Do the observed values fit a certain expected values?

(this can be any expected frequency, e.g. the mean across all categories)

One vs two categorical variables

χ^2 goodness-of-fit

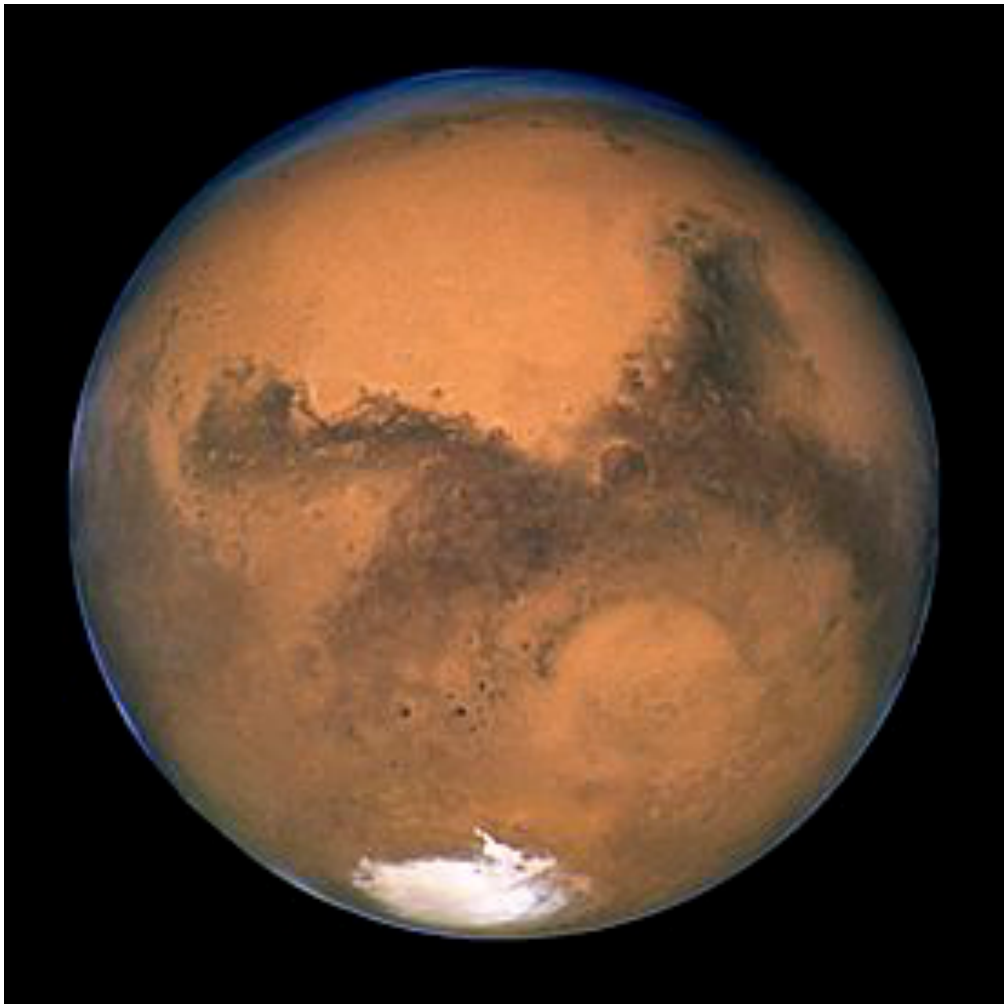
one categorical variable. Do the observed values fit a certain expected values?

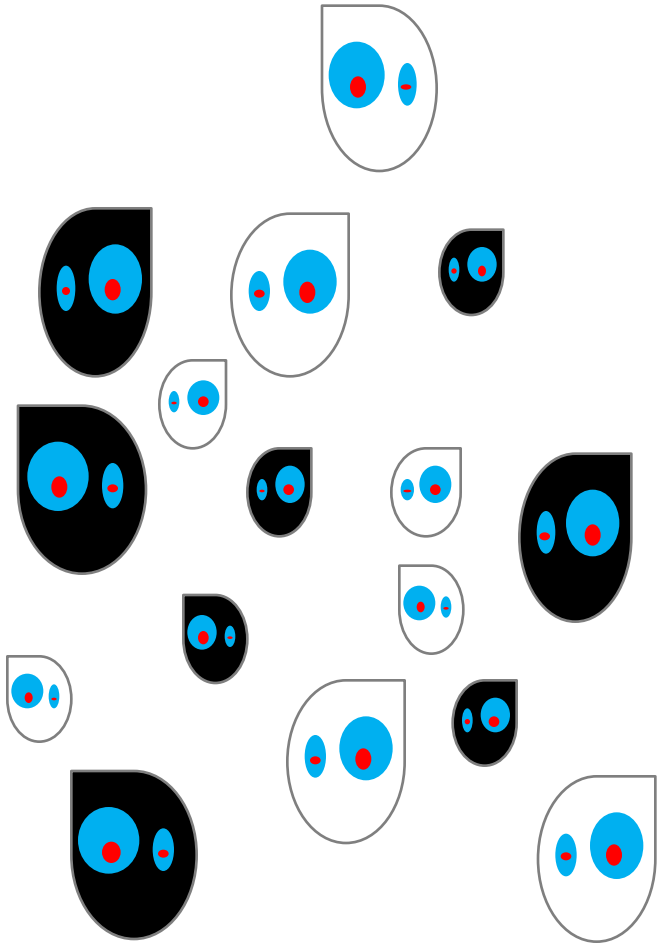
(this can be any expected frequency, e.g. the mean across all categories)

Contingency analyses

test whether observed value fit expected assuming two categorical variables are *independent*

Example





www.wikipedia.com

Example

Far in the future a new life form gets described on Mars; they come in a small and big variety and are black or white.

23rd century BIOL 300 students are interested to know if these two characteristic are independent or not.

A field trip goes to Mars and a sample is taken.

Student 1 looks at the colour and notes: 20 black and 20 white

Student 2 look at size and notes: 20 small and 20 big.

We know how to test whether statistic deviates from a certain expected value (binomial test or χ^2 goodness-of-fit test)

Example

Student 3 notes that both measurements have been taken from the same individuals and wonder whether there is a pattern between colour and size. In other words: *are size and colour independent in these organisms?*

H_0: size and colour are independent of one another

H_A: size and colour are **not** independent of one another

Example

Observed

Size: 20 small and 20 big

Colour: 20 black and 20 white

What is the probability to get

Pr[small] =

Pr[big] = 1 - Pr[small]

Example

Observed

Size: 20 small and 20 big

Colour: 20 black and 20 white

What is the probability to get

Pr[small] = 0.5

Pr[big] = 1 - Pr[small]

Pr[black] =

Pr[white] = 1 - Pr[black]

Example

Assuming samples are independent*, what is thee

Pr[small AND black] =

Pr[large AND black] =

Pr[small AND white] =

Pr[large AND white] =

sum =

*independent + AND = multiplication rule

Example

Assuming samples are independent*, what is the

Pr[small AND black] = 0.5 x 0.5 = 0.25

20/40 = 0.5 20/40 = 0.5

Exp.	black	white	Totals
small	0.5 x 0.5 = 0.25	…	20
big	….	….	20
Totals	20	20	40

20/40 = 0.5

20/40 = 0.5

*independent + AND = multiplication rule

Example

From expected probability to expected frequency = x sample size (40)

20/40 = 0.5 20/40 = 0.5

	Exp.	black	white	Totals
20/40 = 0.5	small	0.25 x 40 = 10	10	20
20/40 = 0.5	big	10	10	20
	Totals	20	20	40

*independent + AND = multiplication rule

What if they are independent?

$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

$\chi^2 = (11 - 10)^2/10 + (10 - 10)^2/10 + (9 - 10)^2/10 + (10 - 10)^2/10 = 0.2$

d.f. = (r - 1)(c - 1)= (2 - 1)(2 - 1) = 1

$\chi^2_{0.05,1} = 3.84$

0.2 < 3.84 do not reject H_0

Obs.	black	white	Totals
small	11 10	9 10	20
big	10 10	10 10	20
Totals	20	20	40

What if they observed are more divergent?

$\chi^2 = (1 - 10)^2/10 + (18 - 10)^2/10 + (19 - 10)^2/10 + (8 - 10)^2/10 = 25.6$

d.f. = (r - 1)(c - 1)= (2 - 1)(2 - 1) = 1

$\chi^2_{0.05,1} = 3.84$

29 > 3.84 reject H_0

Obs.	black	white	Totals
small	2 10	18 10	20
big	18 10	2 10	20
Totals	20	20	40

Problem?

Example 2

Size: 21 small and 30 big (n = 51)

Colour: 15 black and 36 white

What is the probability under H_0 to get

Pr[small] =

Pr[big] = 1 - Pr[small]

Pr[black] =

Pr[white] = 1 - Pr[black]

Expected probabilities for each combination.

Multiply with sample size

15/51 = 0.29 36/51 =0.71

Exp.	black	white	Totals
small	0.41 x 0.29	0.41 x...	21
big	..	..	30
Totals	15	36	51

21/51 = 0.41

1 – Pr[small]

Expected probabilities for each combination.

Multiply with sample size

15/51 = 0.29 36/51 =0.71

Exp.	black	white	Totals
small (21/51 = 0.41)	0.41 x 0.29 = 6.2	14.8	21
big (30/51 = 0.59)	8.8	21.2	30
Totals	15	36	51

Expected frequencies

$$\chi^2 = \sum_i \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

χ^2=0.264

d.f. = (r - 1)(c - 1)= (2 - 1)(2 - 1) = 1

$\chi^2_{0.05,1}$ = 3.84

H_0 …???? .. rejected

15/51 = 0.29 36/51 =0.71

Obs.	black	white	Totals
small	7 6.2	14 14.8	21
big	8 8.8	22 21.2	30
Totals	15	36	51

21/51 = 0.41

30/51 = 0.59

Expected frequencies for a different case…

$\chi^2 = 6.80$

d.f. = (r - 1)(c - 1)= (2 - 1)(2 - 1) = 1

$\chi^2_{0.05,1} = 3.84$

H_0 …???? .. rejected

15/51 = 0.29 36/51 =0.71

Obs.	black	white	Totals
small	2 6.2	19 14.8	21
big	13 8.8	17 21.2	30
Totals	15	36	51

21/51 = 0.41

30/51 = 0.59

Fisher's exact test

- For 2 x 2 contingency analysis
- Does not make assumptions about the size of expectations
- JMP (or other programs) will do it, but cumbersome to do by hand

Example

Some fish can develop into either sex depending on social circumstances. Does the sex that juvenile gobies develop into depend on the sex of the adult fish they were raised with?

Obs.	Became Male	Became Female	Totals
Raised with Male	1	11	12
Raised with Female	6	4	10
Totals	7	15	22

Exp.	Became Male	Became Female	Totals
Raised with Male	12/22 * 7/22 * 22 = **3.8**		12
Raised with Female			10
Totals	7	15	22

Example

Some fish can develop into either sex depending on social circumstances. Does the sex that juvenile gobies develop into depend on the sex of the adult fish they were raised with?

Obs.	Became Male	Became Female	Totals
Raised with Male	1	11	12
Raised with Female	6	4	10
Totals	7	15	22

Exp.	Became Male	Became Female	Totals
Raised with Male	12/22 * 7/22 * 22 = **3.8**	8.2	12
Raised with Female	3.2	6.8	10
Totals	7	15	22

df = (r-1)(c-1) = (2-1)(2-1) = **1**

Example

Some fish can develop into either sex depending on social circumstances. Does the sex that juvenile gobies develop into depend on the sex of the adult fish they were raised with?

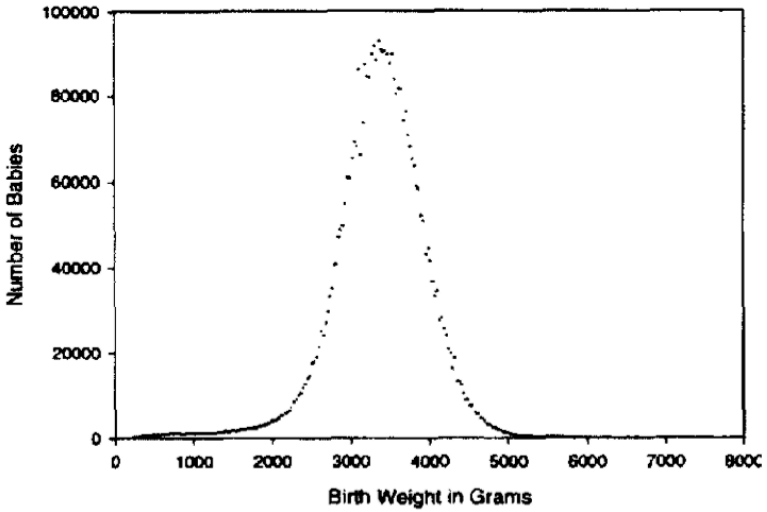
Obs.	Became Male	Became Female	Totals
Raised with Male	1	11	12
Raised with Female	6	4	10
Totals	7	15	22

Exp.	Became Male	Became Female	Totals
Raised with Male	3.8	8.2	12
Raised with Female	3.2	6.8	10
Totals	7	15	22

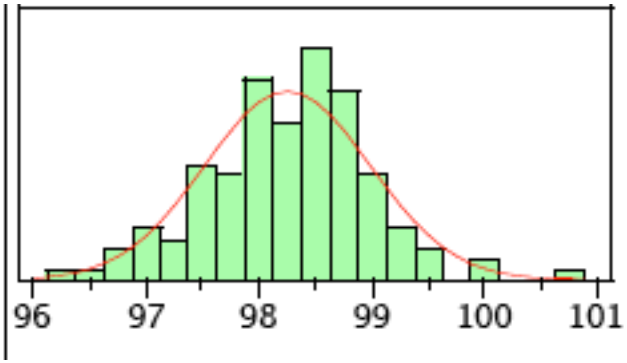
Any problems using χ^2 ?

The normal distribution

The normal distribution is very common in nature

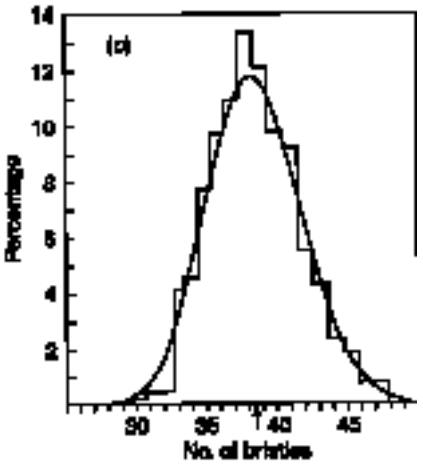


Human birth weight



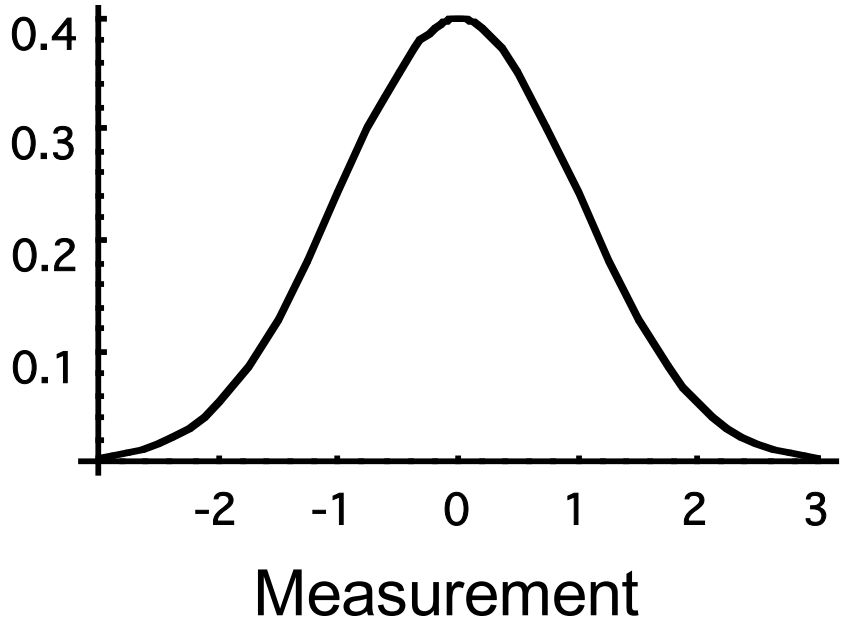
Human body temperature

Number of bristles on a *Drosophila abdomen*

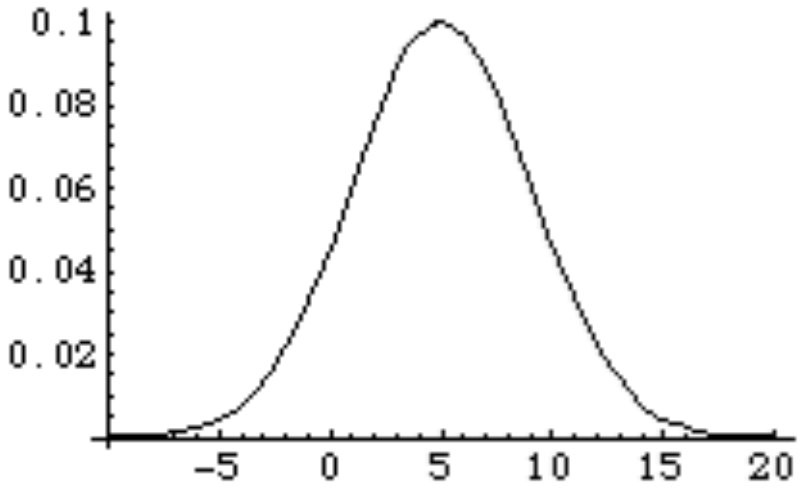


Normal distribution

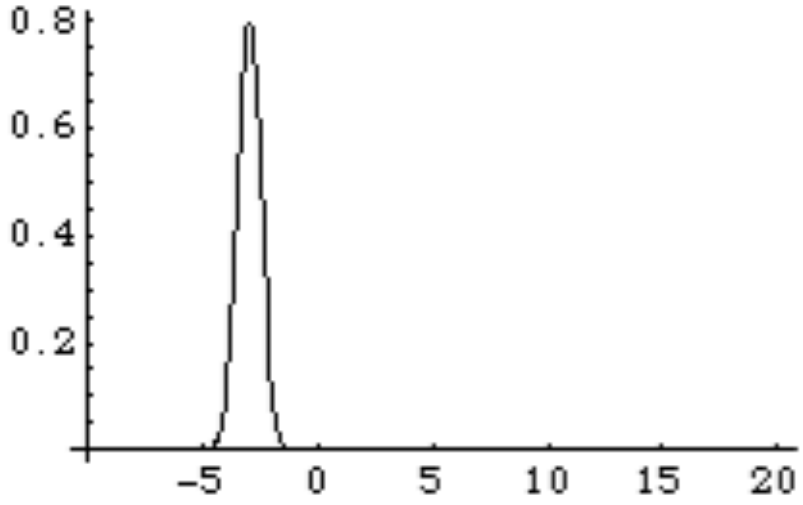
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



A normal distribution is fully described by its mean and standard deviation

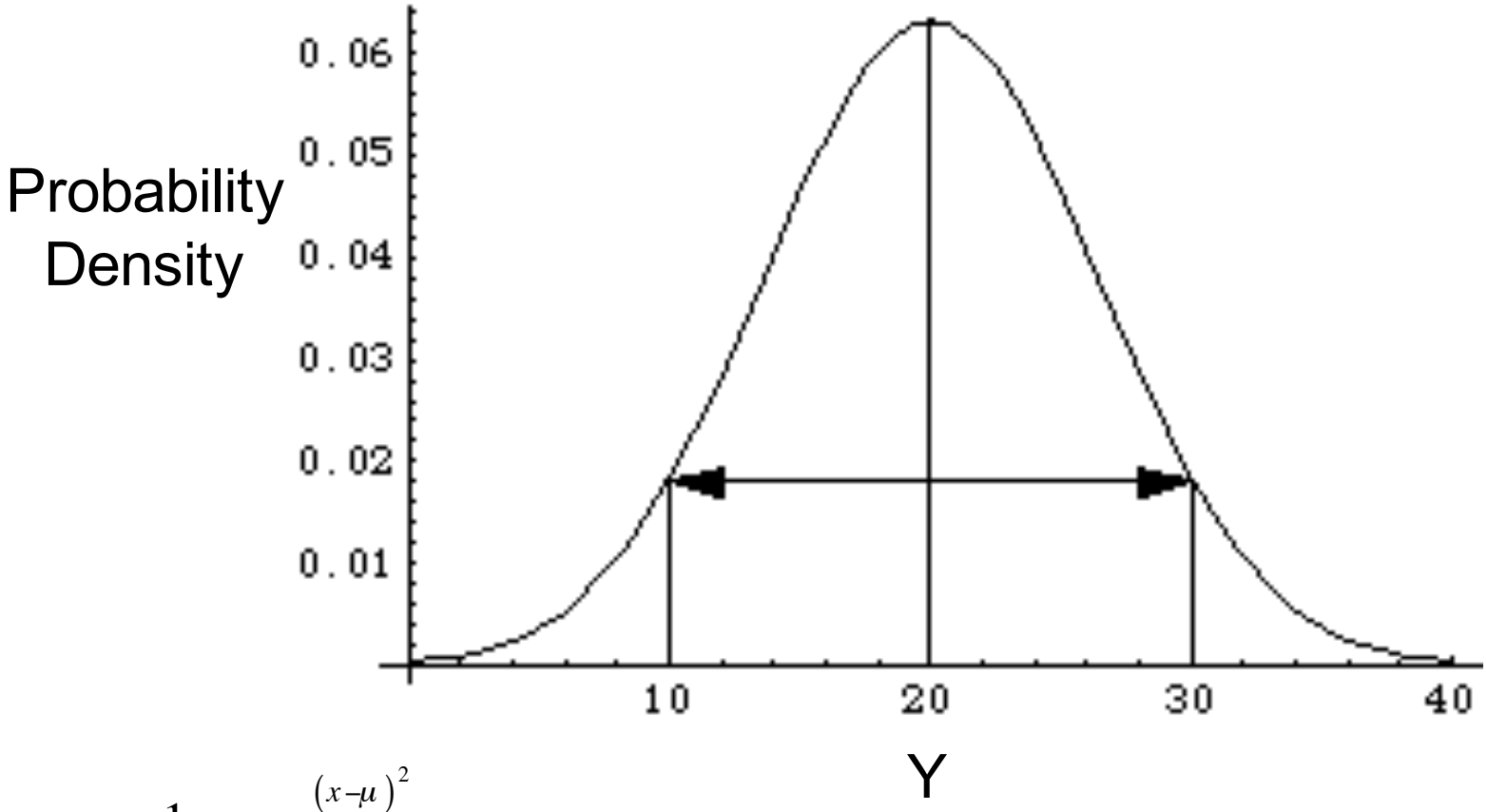


$\mu = 5; \sigma = 4$



$\mu = -3; \sigma = 1/2$

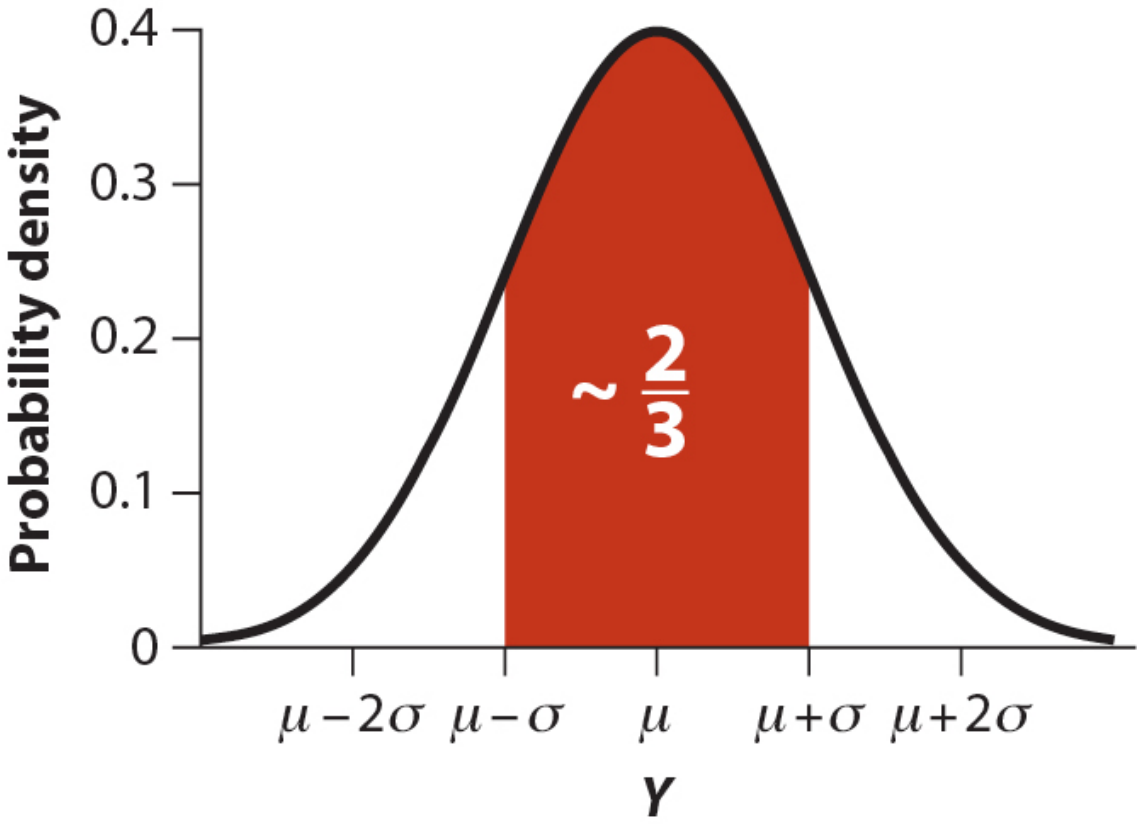
A normal distribution is symmetric around its mean



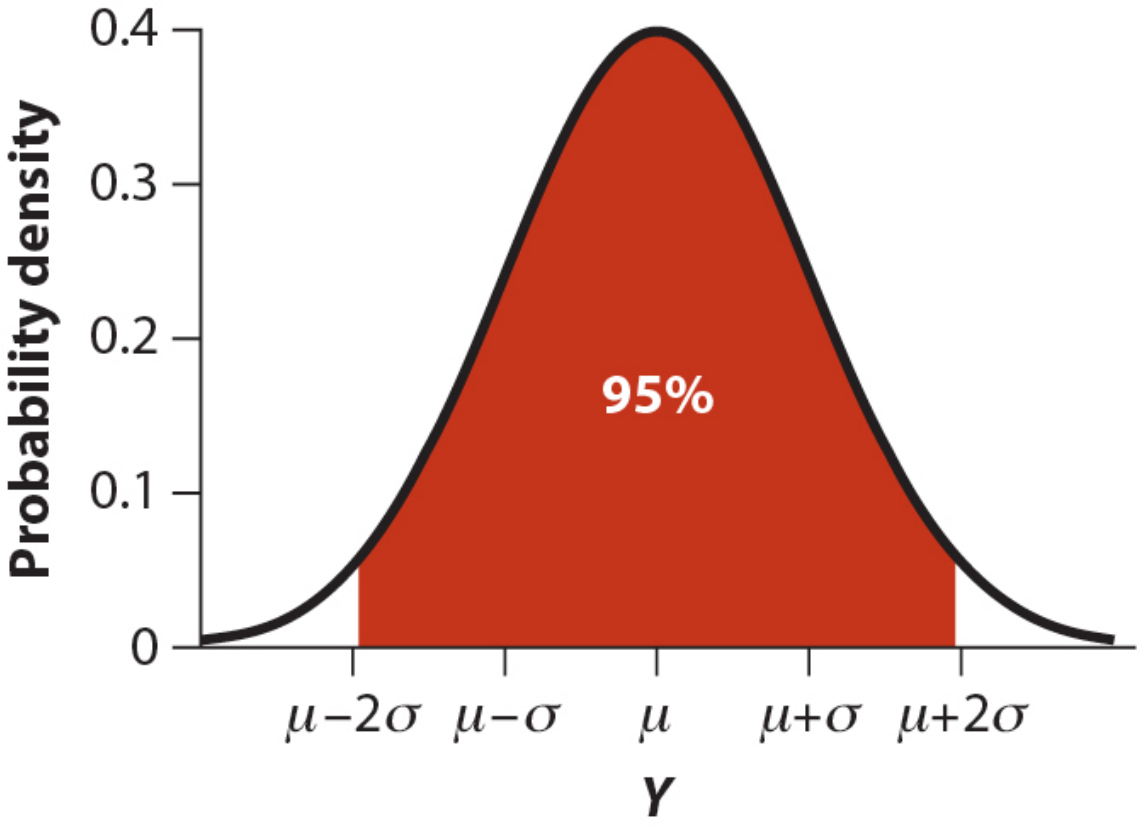
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

With a normal distribution, the mean, median and mode are all the same.

About 2/3 of random draws from a normal distribution are within one standard deviation of the mean



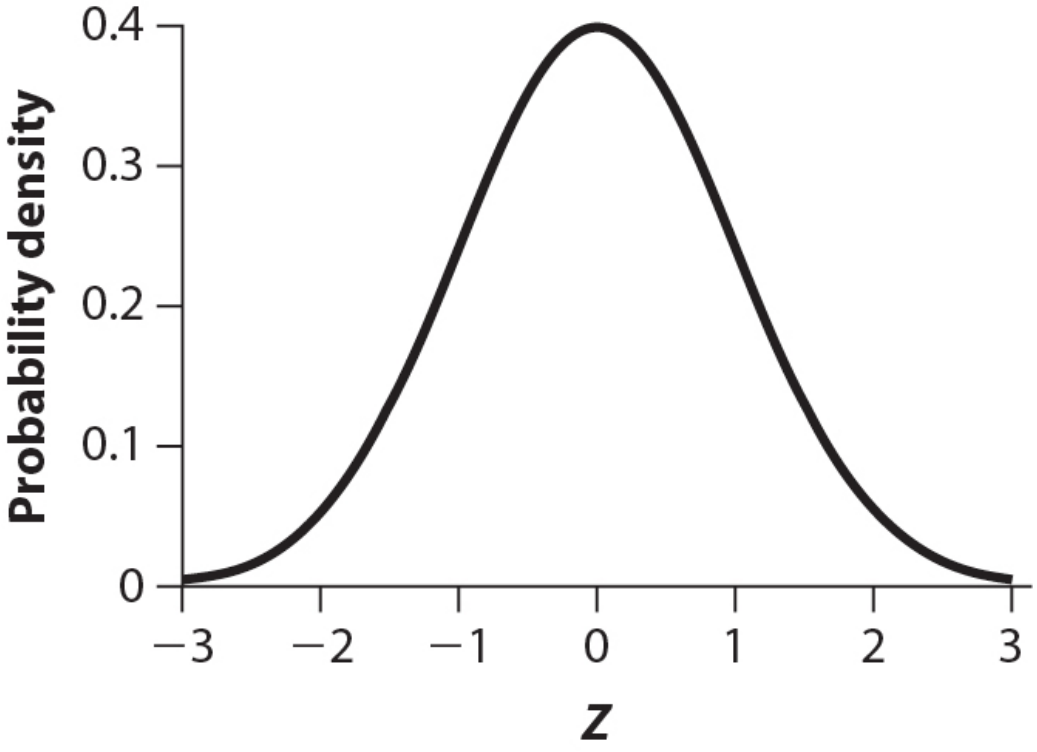
About 95% of random draws from a normal distribution are within two standard deviations of the mean



(Really, it's 1.96 SD.)

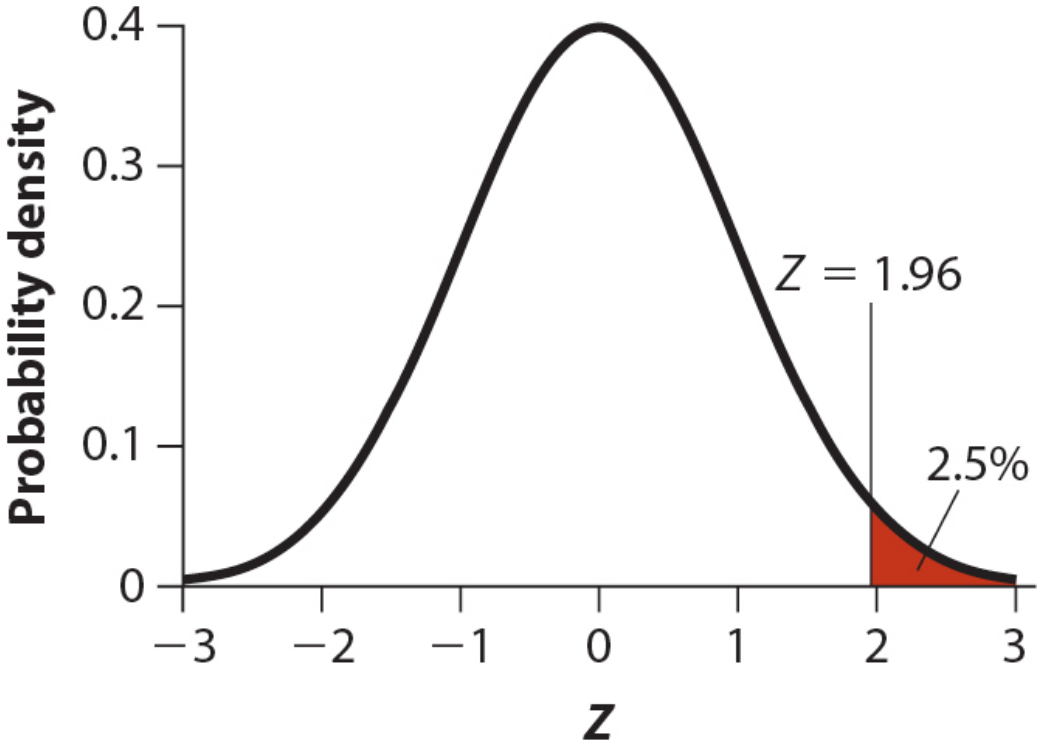
Standard normal distribution

- Mean is zero. ($\mu = 0$)
- Standard deviation is one. ($\sigma = 1$)



Standard normal table

- Gives the probability of getting a random draw from a standard normal distribution greater than a given value



Standard normal table: $Z = 1.96$

First two digits of *a.bc*	Second digit after decimal (*c*)									
	0	1	2	3	4	5	6	7	8	9
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426

Standard normal is symmetric, so...

- $\Pr[Z > x] = \Pr[Z < -x]$
- $\Pr[Z < x] = 1 - \Pr[Z > x]$

What about other normal distributions?

- All normal distributions are shaped alike, just with different means and variances
- Any normal distribution can be converted to a standard normal distribution, by

$$Z = \frac{Y - \mu}{\sigma}$$

Z is called a "*standard normal deviate*."

$$Z = \frac{Y - \mu}{\sigma}$$

Z tells us how many standard deviations *Y* is from the mean

The probability of getting a value greater than Y is the same as the probability of getting a value greater than *Z* from a standard normal distribution.

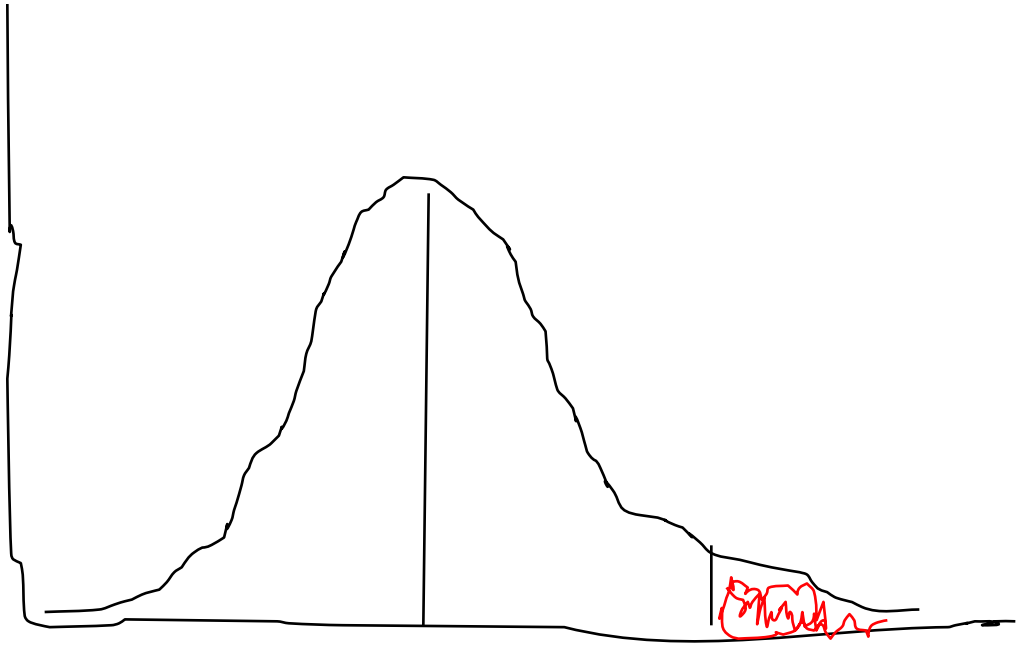
Example: British spies

MI5 says a man has to be shorter than 180.3 cm tall to be a spy.

Mean height of British men is 177.0cm, with standard deviation 7.1cm, with a normal distribution.

What proportion of British men are excluded from a career as a spy by this height criteria?

Draw a rough sketch of the question



μ = 177.0cm
σ = 7.1cm
Y = 180.3

Pr[*height* > *180.3*]

$$Z = \frac{Y - \mu}{\sigma}$$

$$Z = \frac{180.3 - 177.0}{7.1}$$

$$Z = 0.46$$

Part of the standard normal table

	x.x0	x.x1	x.x2	.x3	x.x4	x.x5	x.x6	x.x7	x.x8	x.x9
0.0	0.5	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.4721	0.46812	0.46414
0.1	0.46017	0.4562	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.3707	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.3409	0.33724	0.3336	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.2946	0.29116	0.28774	0.28434	0.28096	0.2776

Pr[Z > 0.46] = 0.32276,

so Pr[*height* > 180.3] = 0.32276

Example

NASA excludes anyone under 157.5 cm and over 190.5 cm in height from being an astronaut pilot. What proportion of American men are excluded from being an astronaut pilot?

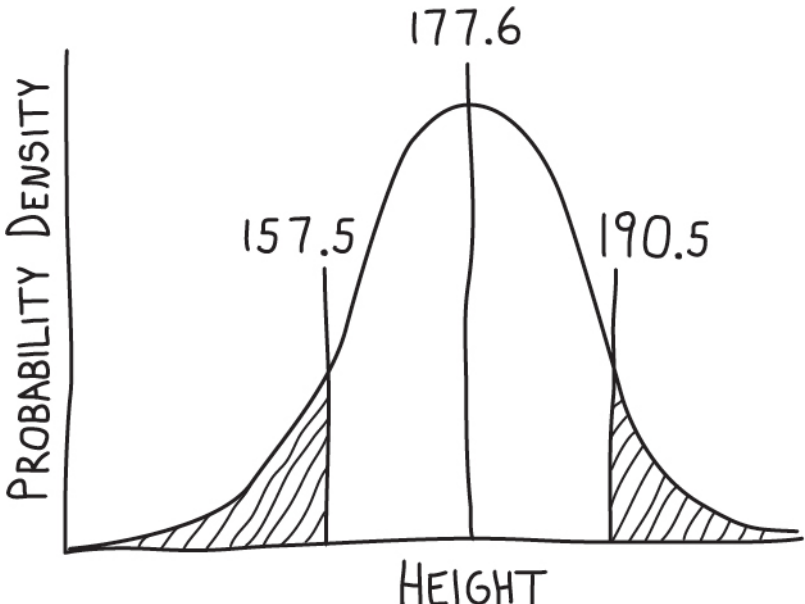
Height is normally distributed with mean 177.6 cm and standard deviation of 9.7 cm.

Example

NASA excludes anyone under 157.5 cm and over 190.5 cm in height from being an astronaut pilot. What proportion of American men are excluded from being an astronaut pilot?

Height is normally distributed with mean 177.6 cm and standard deviation of 9.7 cm.

Pr[Height < 157.5 or Height > 190.5]



μ = 177.6 cm
σ = 9.7 cm
Y = 190.5 cm

Pr[*height* > *190.5*]

$$Z = \frac{Y - \mu}{\sigma}$$

$$Z = \frac{190.5 - 177.6}{9.7}$$

$$Z = 1.33$$

Part of the standard normal table

First two digits of *a.bc*	Second digit after decimal (*c*)									
	0	1	2	3	4	5	6	7	8	9
0.0	0.5	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
			0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494
									0.00391	0.00379

Pr[$Z > 1.33$] = 0.09176,

so Pr[*height* > 190.5] = 0.09176

μ = 177.6 cm
σ = 9.7 cm
Y = 157.5 cm

Pr[*height* < *157.5*]

$$Z = \frac{Y - \mu}{\sigma}$$

$$Z = \frac{157.5 - 177.6}{9.7}$$

$$Z = -2.07$$

Part of the standard normal table

First two digits of *a.bc*	Second digit after decimal (*c*)									
	0	1	2	3	4	5	6	7	8	9
0.0	0.5	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
			0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
						0.00402	0.00391	0.00379	0.00368	0.00357

Where are the negative values?

Part of the standard normal table

First two digits of *a.bc*	Second digit after decimal (*c*)									
	0	1	2	3	4	5	6	7	8	9
0.0	0.5	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414
0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
			0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
						0.00402	0.00391	0.00379	0.00368	0.00357

Pr[$Z < -2.07$] = Pr[$Z > 2.07$] =0.01923 ,
so Pr[*height* < 157.5] = 0.01923

Example

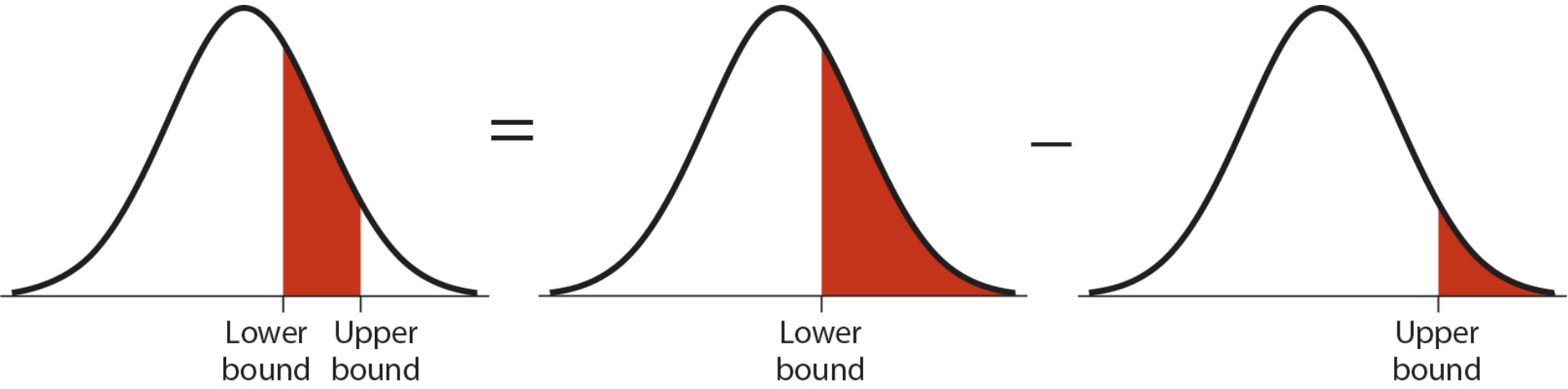
NASA excludes anyone under 157.5 cm and over 190.5 cm in height from being an astronaut pilot. What proportion of American men are excluded from being an astronaut pilot?

Height is normally distributed with mean 177.6 cm and standard deviation of 9.7 cm.

Pr[Height < 157.5 or Height > 190.5] = 0.01923 + 0.09176 = 0.11099

11.1% of American males are excluded from being an astronaut by their height.

Probability Z lies between two values



Pr[lower bound < Z < upper bound] = Pr[Z > lower bound] – Pr[Z > upper bound]

Sample means are normally distributed

(If the variable itself is normally distributed.)

- The mean of the sample means is μ.
- The standard deviation of the sample means is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

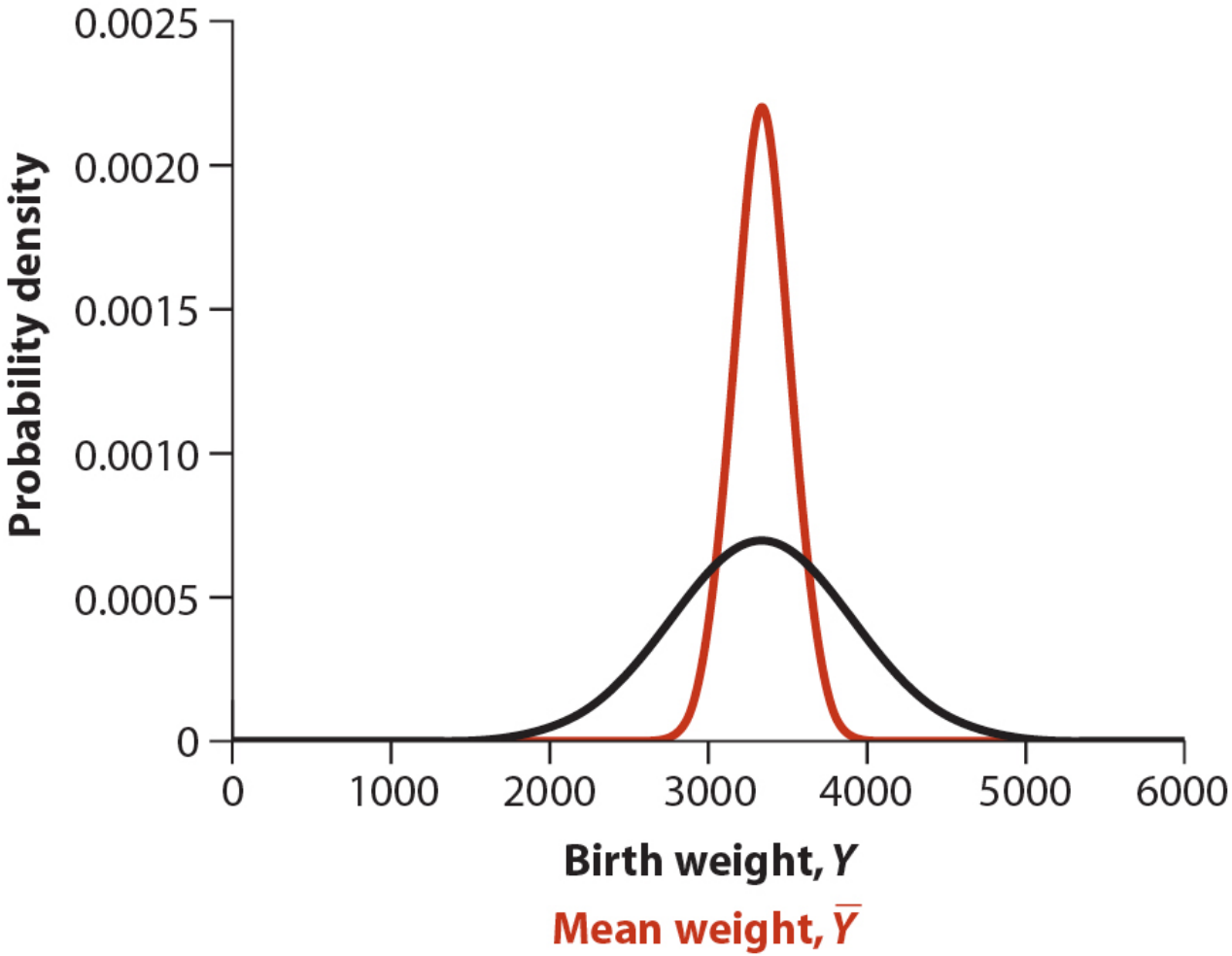
Standard error

The standard error of an estimate of a mean is the standard deviation of the distribution of sample means

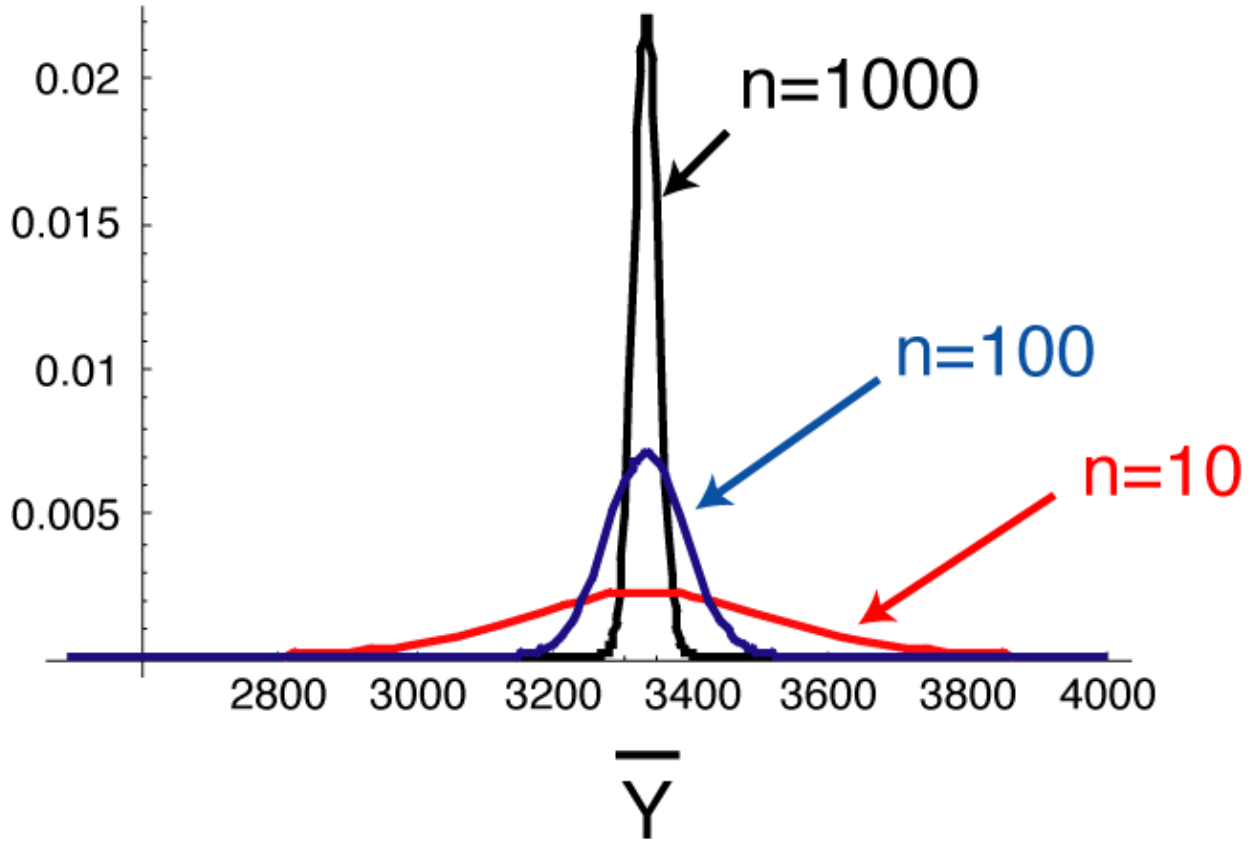
$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

We can approximate this by $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$

Distribution of means of samples with n =10



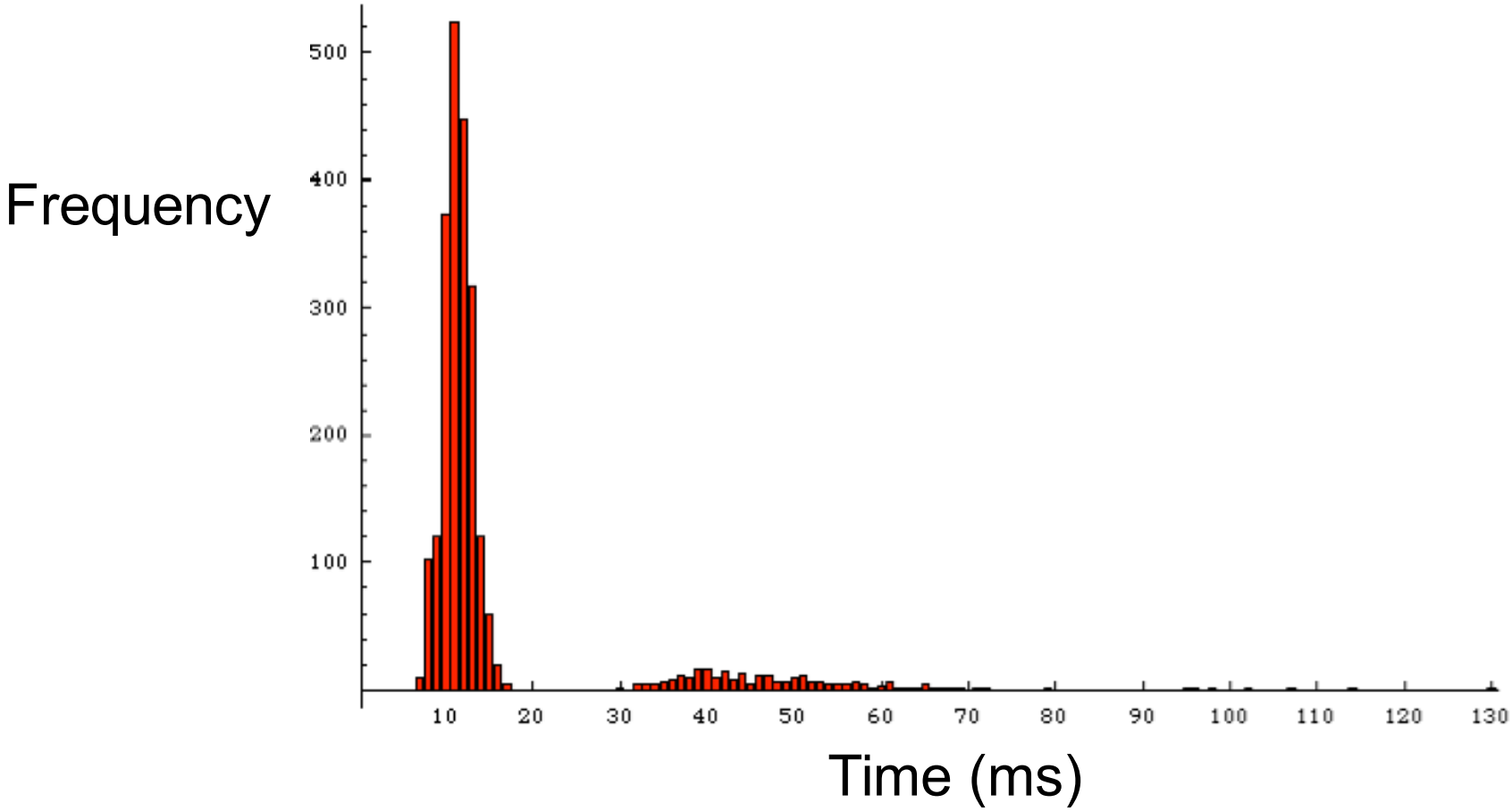
Larger samples equal smaller standard errors

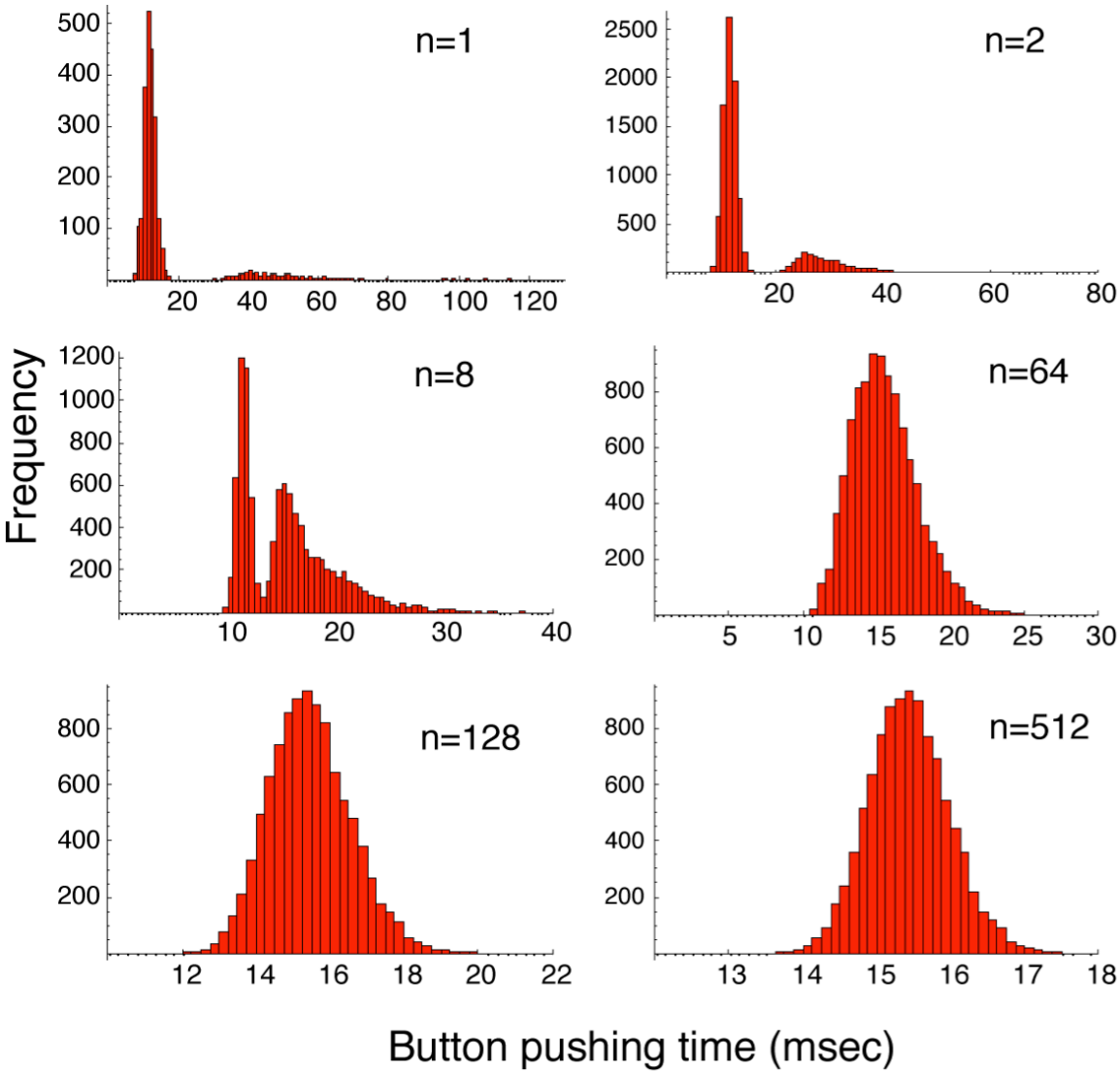


Central limit theorem

The sum or mean of a large number of measurements randomly sampled from *any* population is approximately normally distributed.

Button pushing times

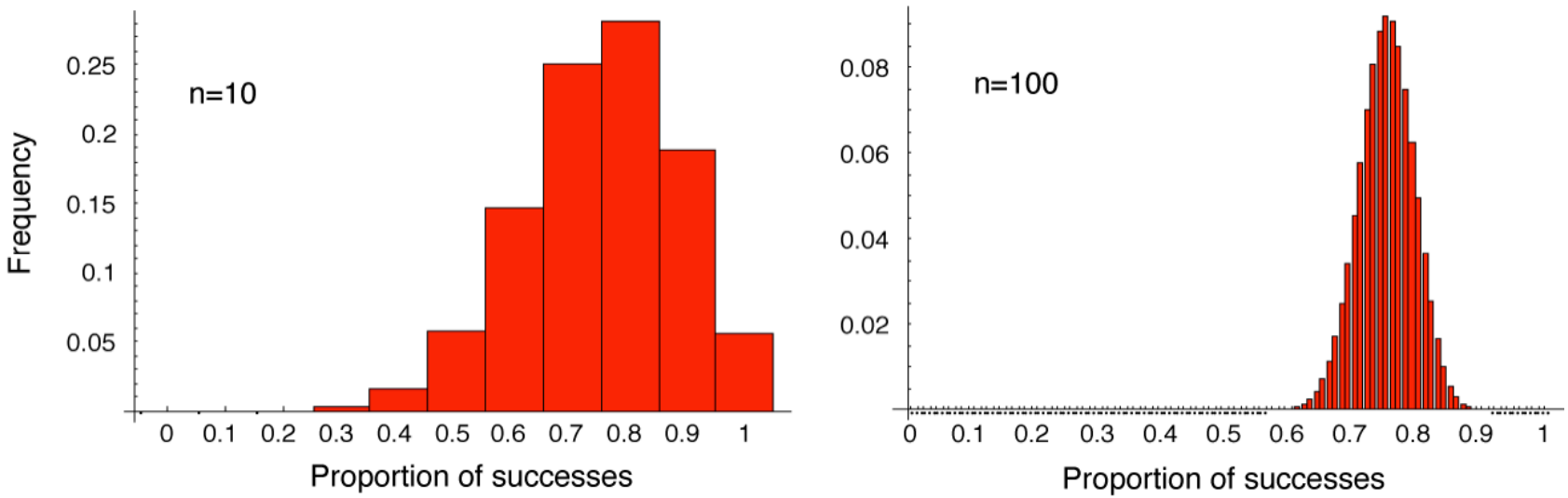



n=1
n=2
n=8
n=64
n=128
n=512
Frequency
Button pushing time (msec)

Normal approximation to the binomial distribution

The binomial distribution, when number of trials n is large and probability of success p is not close to 0 or 1, is approximated by a normal distribution having mean np and standard deviation $\sqrt{np(1-p)}$.

The binomial distribution approaches a normal distribution as sample size gets larger



This is an example of the Central Limit Theorem in action.

Normal approximation to the binomial distribution

$$\Pr\left[\text{number of successes} \geq X\right] = \Pr\left[Z > \frac{X - np}{\sqrt{np(1-p)}}\right]$$

Read more in the text...

When can we use it?

Rule of thumb

- *np* and *n(1-p)* both should be > 5

If you only want to know whether p > or < 0.05

- *np* should be larger if you need higher precision for p value

Normal approximation to the binomial distribution

$$\Pr[X \geqq Observed] \sim \Pr[Z > \frac{(Observed - np)}{\sqrt{np(1-p)}}]$$

Observed is particular value of X observed in the data

Remember that under H_0

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

Normal approximation to the binomial distribution

$$\Pr[X \geq Observed] \sim \Pr[Z > \frac{(Observed\ - np)}{\sqrt{np(1-p)}}]$$

With correction for continuity (binomial is discrete, normal continuous)

$$\Pr[X \geq Observed] \sim \Pr[Z > \frac{(Observed\ - \frac{1}{2} - np)}{\sqrt{np(1-p)}}]$$

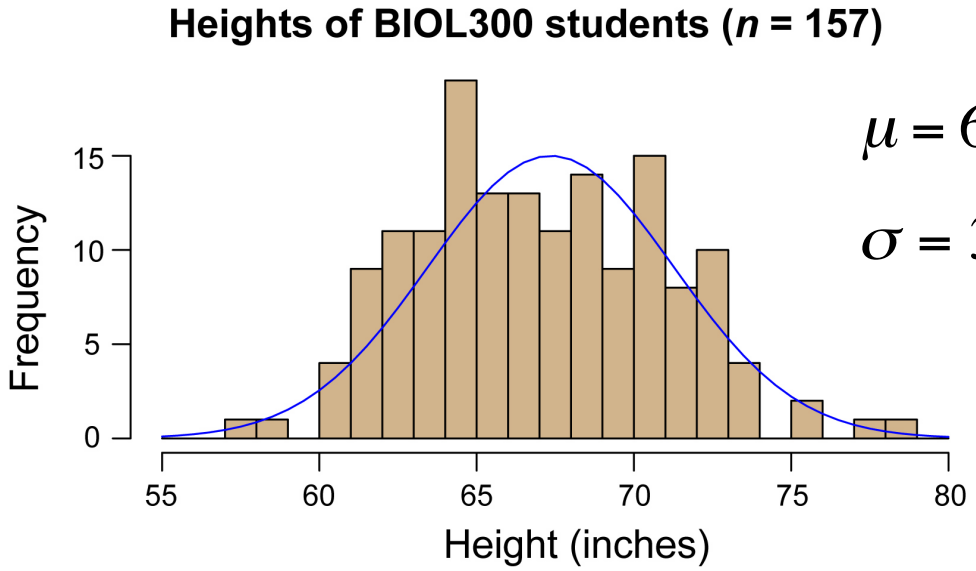
Normal approximation to the binomial distribution

With correction for continuity

$$\Pr[X \geq Observed] \sim \Pr[Z > \frac{(Observed\ - \frac{1}{2} - np)}{\sqrt{np(1-p)}}]$$

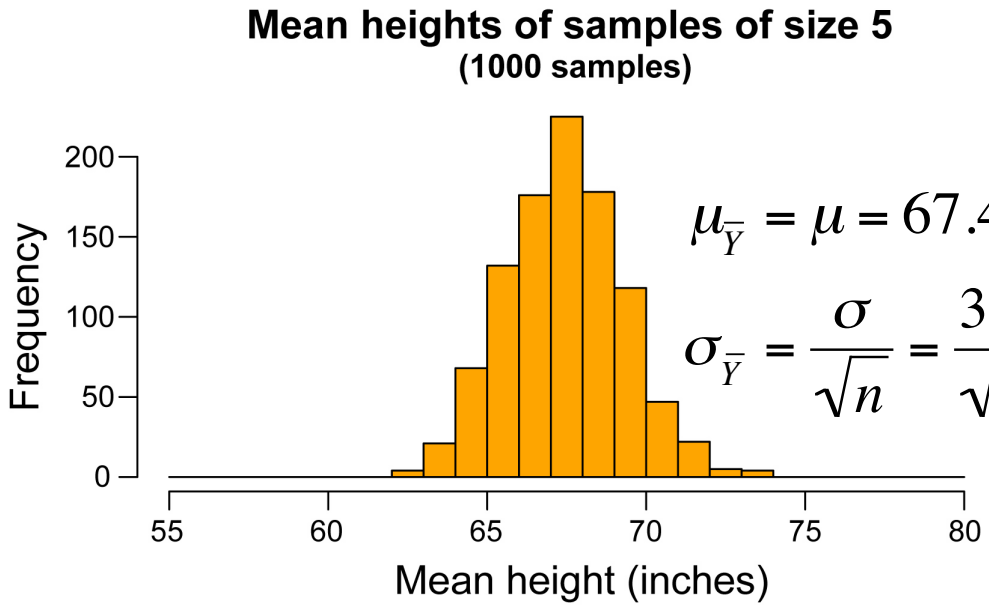
$$\Pr[X \leq Observed] \sim \Pr[Z > \frac{(Observed + \frac{1}{2} - np)}{\sqrt{np(1-p)}}]$$

Inference from a normal population



$$\mu = 67.4$$

$$\sigma = 3.9$$



$$\mu_{\bar{Y}} = \mu = 67.4$$

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.9}{\sqrt{5}} = 1.7$$

$\bar{Y}$ is normally distributed whenever:

Y is normally distributed

or

n is large

Inference about means

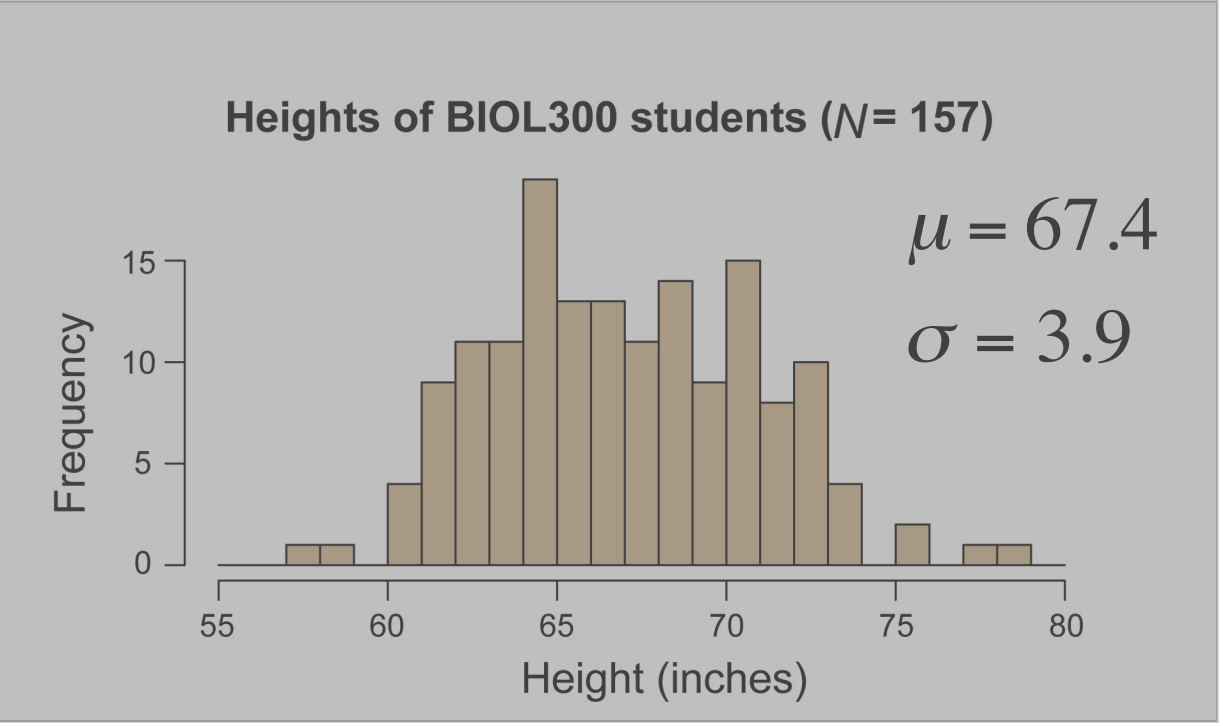
Because $\bar{Y}$ is normally distributed, we can convert its distribution to a standard normal distribution:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

This would give a probability distribution of the difference between a sample mean and the population mean.

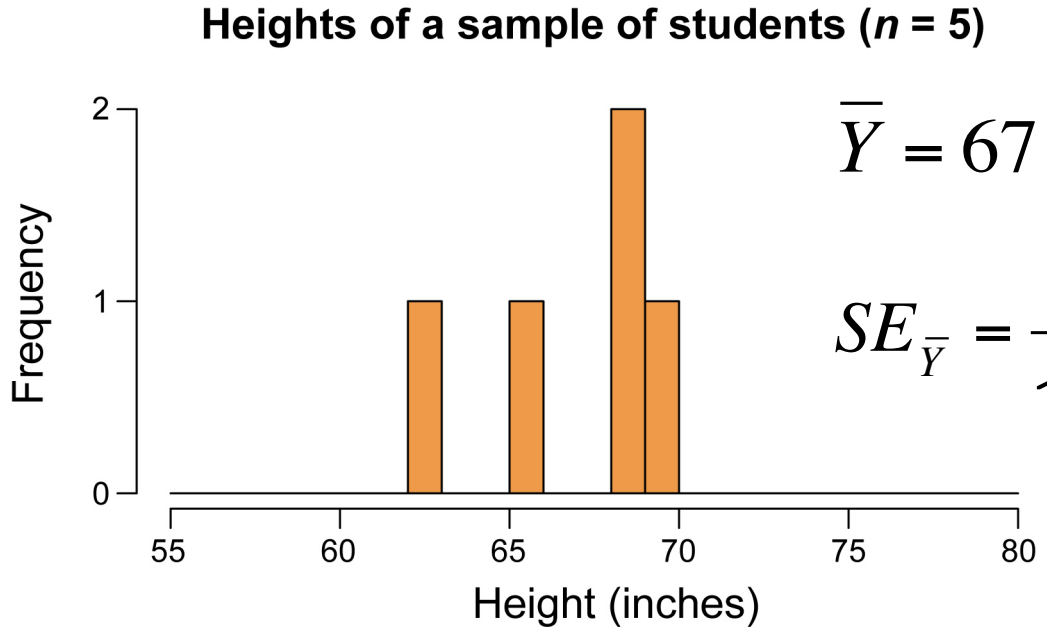
But... we don't know σ . . .

However, we do know *s*, the standard deviation of our sample. We can use that as an estimate of σ.



In most cases, we don't know the real population distribution.

We only have a sample.



$$\bar{Y} = 67.1 \qquad s = 3.1$$

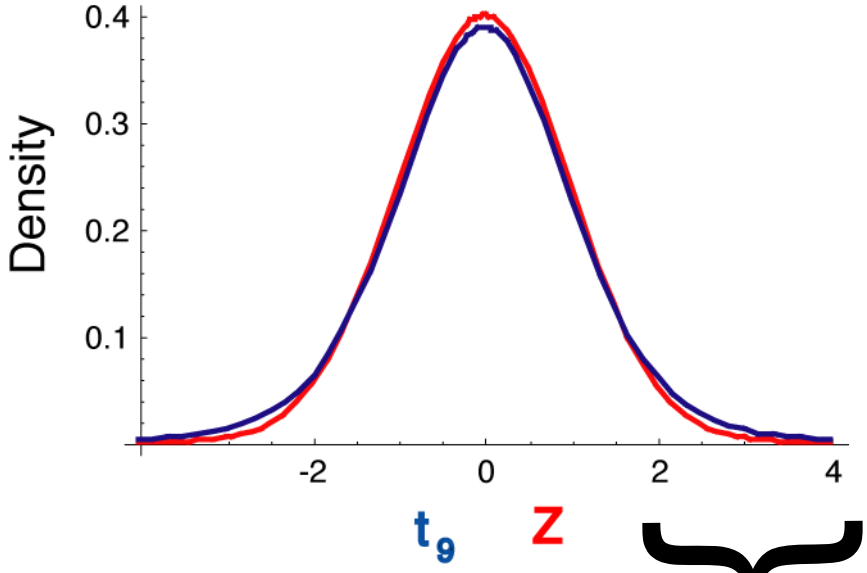
$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{5}} = 1.4$$

We use this as an estimate of $\sigma_{\bar{Y}}$

A good approximation to the standard normal is then:

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

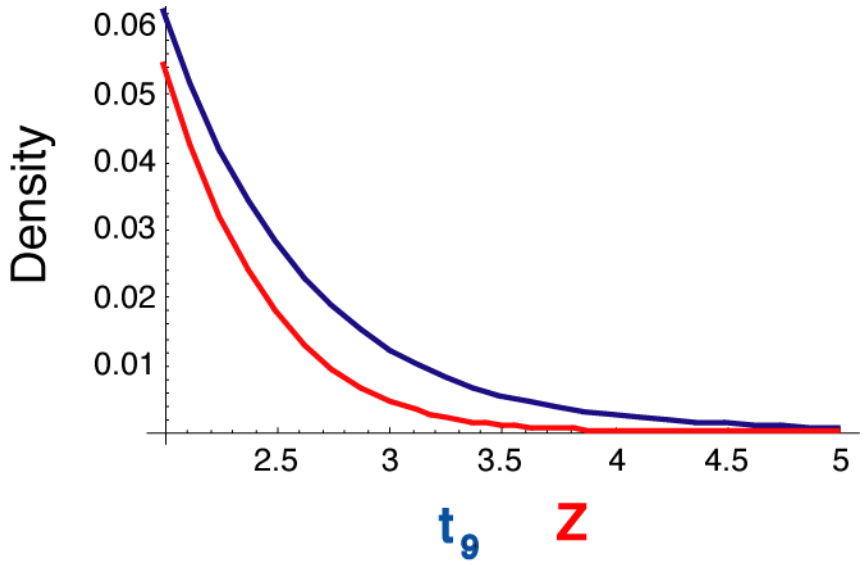
t has a Student's *t* distribution



Discovered by William Gossett, of the Guinness Brewing Company

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$



Student a.k.a William *Gossett*

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

wikipedia

Degrees of freedom

$$df = n - 1$$

Critical values in the t-distribution

$t_{\alpha(\#\ tails),df}$

$t_{0.05(2),4} = \pm 2.78$

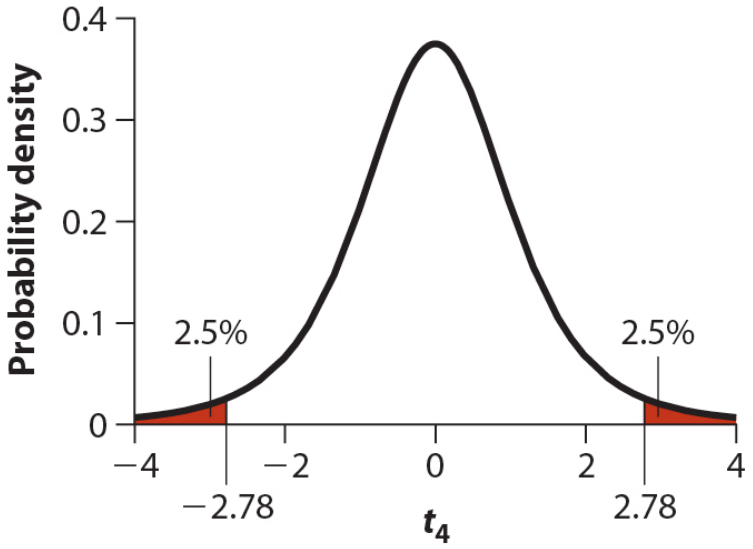


Table C: Student's *t* distribution:

	α(2):	0.2	0.10	0.05	0.02	0.01	0.001	0.0001
df	α(1):	0.1	0.05	0.025	0.01	0.005	0.0005	0.00005
1		3.08	6.31	12.71	31.82	63.66	636.62	6366.20
2		1.89	2.92	4.30	6.96	9.92	31.60	99.99
3		1.64	2.35	3.18	4.54	5.84	12.92	28.00
4		1.53	2.13	2.78	3.75	4.60	8.61	15.54
5		1.48	2.02	2.57	3.36	4.03	6.87	11.18
6		1.44	1.94	2.45	3.14	3.71	5.96	9.08
7		1.41	1.89	2.36	3.00	3.50	5.41	7.88
8		1.40	1.86	2.31	2.90	3.36	5.04	7.12
9		1.38	1.83	2.26	2.82	3.25	4.78	6.59

We use the *t*-distribution to calculate an exact confidence interval of the mean

$$-t_{\alpha(2),df} < \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} < t_{\alpha(2),df}$$

We rearrange the above to generate:

$$\bar{Y} - t_{\alpha(2),df} SE_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} SE_{\bar{Y}}$$

Another way to express this is: $\bar{Y} \pm SE_{\bar{Y}}\, t_{\alpha(2),df}$

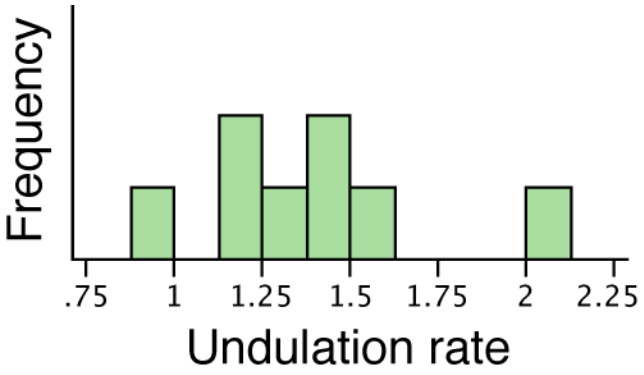
95% confidence interval for a mean

Example:
Paradise flying snakes

Undulation rates (in Hz)

0.9, 1.4, 1.2, 1.2, 1.3, 2.0, 1.4, 1.6

Estimate the mean and standard deviation



$$\bar{Y} = 1.375$$

$$s = 0.324$$

$$n = 8$$

Find the standard error

$$\bar{Y} \pm SE_{\bar{Y}}\, t_{\alpha(2),df}$$

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{0.324}{\sqrt{8}} = 0.115$$

Find the critical value of *t*

$$df = n - 1 = 7$$

$$t_{\alpha(2),df} = t_{0.05(2),7}$$

Table C: Student's *t* distribution

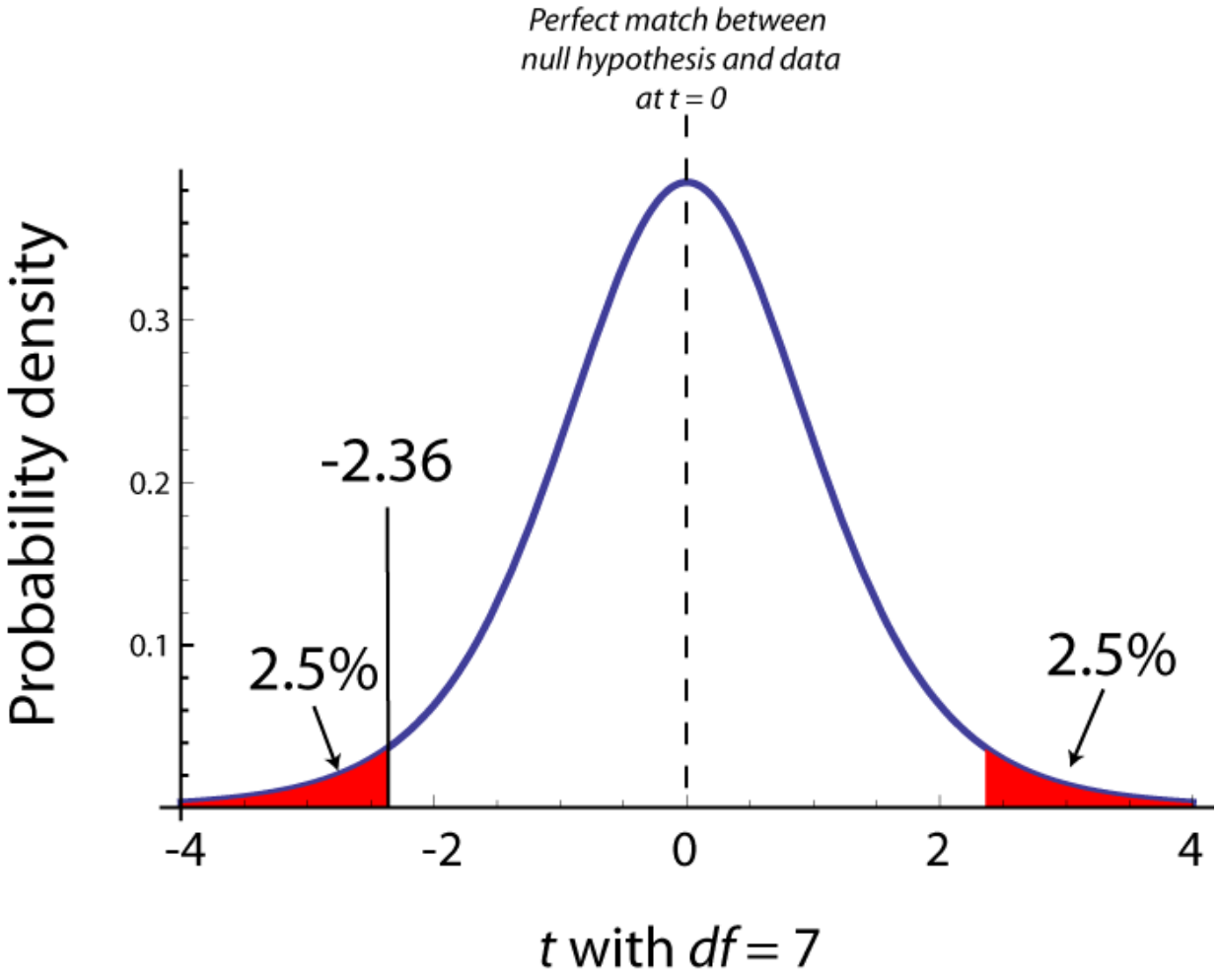
	α(2):	0.2	0.10	0.05	0.02	0.01	0.001	0.0001
df	α(1):	0.1	0.05	0.025	0.01	0.005	0.0005	0.00005
1		3.08	6.31	12.71	31.82	63.66	636.62	6366.20
2		1.89	2.92	4.30	6.96	9.92	31.60	99.99
3		1.64	2.35	3.18	4.54	5.84	12.92	28.00
4		1.53	2.13	2.78	3.75	4.60	8.61	15.54
5		1.48	2.02	2.57	3.36	4.03	6.87	11.18
6		1.44	1.94	2.45	3.14	3.71	5.96	9.08
7		1.41	1.89	2.36	3.00	3.50	5.41	7.88
8		1.40	1.86	2.31	2.90	3.36	5.04	7.12
9		1.38	1.83	2.26	2.82	3.25	4.78	6.59

Find the critical value of *t*

$$df = n - 1 = 7$$

$$t_{\alpha(2),df} = t_{0.05(2),7}$$

$$= 2.36$$



Putting it all together...

$$\bar{Y} \pm SE_{\bar{Y}}\, t_{\alpha(2),df} = 1.375 \pm 0.115 \quad (2.36)$$

$$= 1.375 \pm 0.271$$

$$1.10 < \mu < 1.65$$

(95% confidence interval)

99% confidence interval

$$t_{\alpha(2),df} = t_{0.01(2),7}$$

	α(2):	0.2	0.10	0.05	0.02	0.01	0.001	0.0001
df	α(1):	0.1	0.05	0.025	0.01	0.005	0.0005	0.00005
1		3.08	6.31	12.71	31.82	63.66	636.62	6366.20
2		1.89	2.92	4.30	6.96	9.92	31.60	99.99
3		1.64	2.35	3.18	4.54	5.84	12.92	28.00
4		1.53	2.13	2.78	3.75	4.60	8.61	15.54
5		1.48	2.02	2.57	3.36	4.03	6.87	11.18
6		1.44	1.94	2.45	3.14	3.71	5.96	9.08
7		1.41	1.89	2.36	3.00	3.50	5.41	7.88
8		1.40	1.86	2.31	2.90	3.36	5.04	7.12
9		1.38	1.83	2.26	2.82	3.25	4.78	6.59

99% confidence interval

$$t_{\alpha(2),df} = t_{0.01(2),7} = 3.50$$

$$\bar{Y} \pm SE_{\bar{Y}} \quad t_{\alpha(2),df} = 1.375 \pm 0.115 \quad (3.50)$$

$$= 1.375 \pm 0.403$$

$$0.97 < \mu < 1.78$$

One-sample *t*-test

The *one-sample* t-*test* compares the mean of a random sample from a normal population with the population mean proposed in a null hypothesis.

Test statistic for one-sample *t*-test

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

μ_0 is the mean value proposed by H_0

Hypotheses for one-sample *t*-tests

H_0 : The mean of the population is μ_0.

H_A: The mean of the population is not μ_0.

Example: Human body temperature

H_0 : Mean healthy human body temperature is 98.6ºF.

H_A: Mean healthy human body temperature is not 98.6ºF.

Human body temperature

$$n = 24$$

$$\bar{Y} = 98.28$$

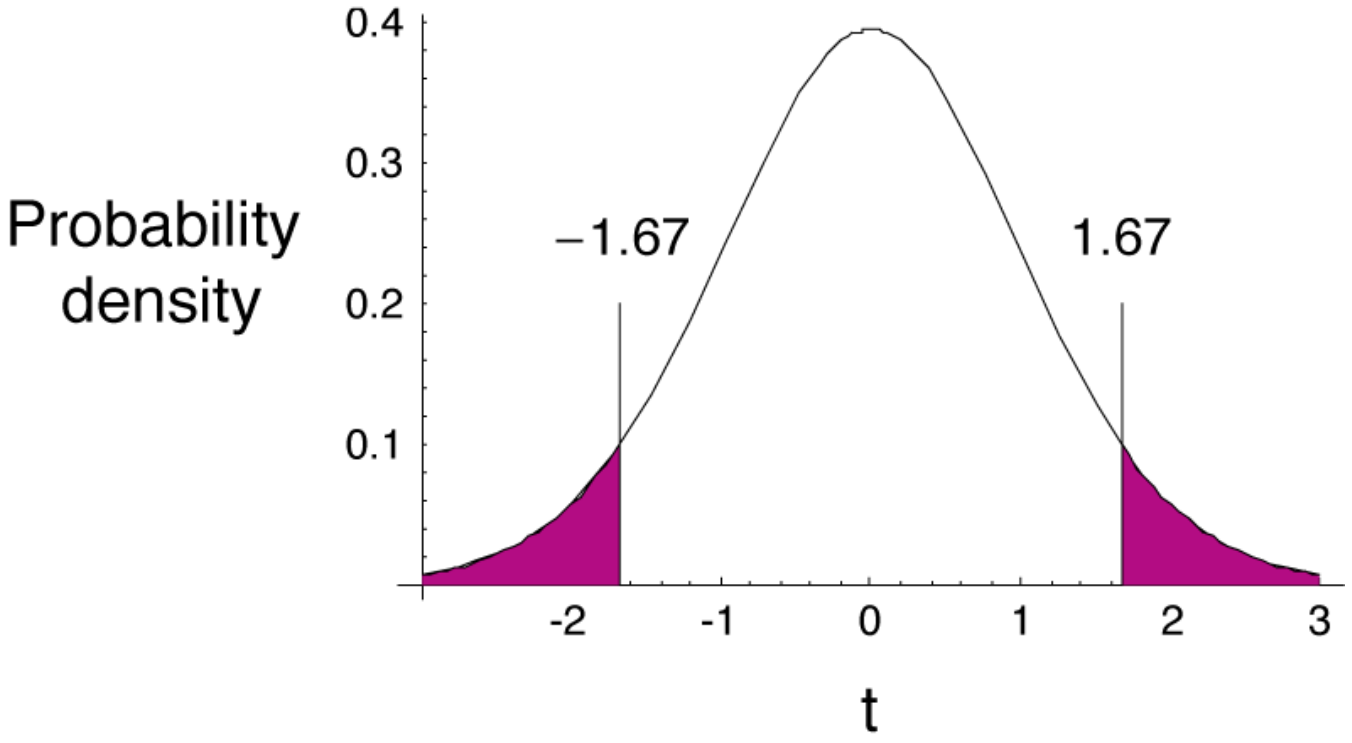
$$s = 0.940$$

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{98.28 - 98.6}{0.940/\sqrt{24}} = -1.67$$

Degrees of freedom

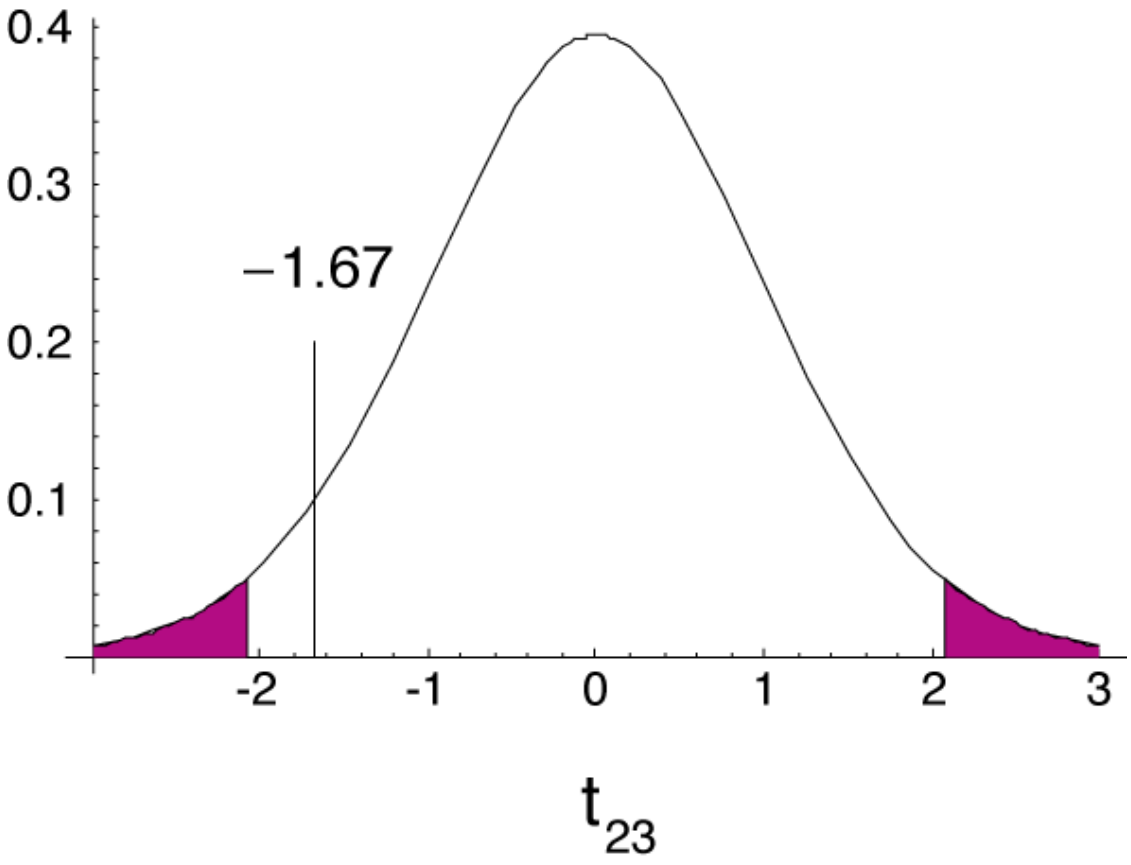
$$df = n - 1 = 23$$

Comparing *t* to its distribution to find the *P*-value



A portion of the *t* table

df	α(1) =0.1 α(2)=0.2	α(1) =0.05 α(2)=0.10	α(1) =0.025 α(2)=0.05	α(1) =0.01 α(2)=0.02	α(1) =0.005 α(2)=0.01
...	...	...	...	...	...
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.5	2.81
24	1.32	1.71	2.06	2.49	2.8
25	1.32	1.71	2.06	2.49	2.79


0.4
0.3
0.2
0.1
−1.67
-2
-1
0
1
2
3
t23

-1.67 is closer to 0 than -2.07, so $P > 0.05$.

With these data, we cannot reject the null hypothesis that the mean human body temperature is 98.6.

Body temperature revisited: $n = 130$

$$n = 130$$

$$\bar{Y} = 98.25$$

$$s = 0.733$$

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{98.25 - 98.6}{0.733/\sqrt{130}} = -5.44$$

Body temperature revisited: $n = 130$

$$t = -5.44$$

$$t_{0.05(2),129} = \pm 1.98$$

t is further out in the tail than the critical value, so we could reject the null hypothesis. Human body temperature is n*ot* 98.6ºF.

One-sample *t*-test: Assumptions

- The variable is normally distributed.
- The sample is a random sample.