Lecture 18: One-Sample Means With the t-Distribution

Chapter 5.3

Goals for Today

- What do we do when n is small?

Sample Size n

We need a large n for two reasons:

1. CLT so sampling distribution of $\overline{x}$ is normal regardless of the true population distribution.
2. ensure s is a good point estimate of σ, so $SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

What if n is small? We're stuck except when:

1. the sample observations are independent
2. the observations are from a population distribution that is normal

the sampling distribution of $\overline{x}$ is nearly normal regardless of n.

Verifying Normality of Population Distribution

When verifying normality for small n, it is important to not only examine the data but also think about where the data come from. For example, ask:

- Would I expect this distribution to be symmetric?
- Am I confident that outliers are rare?

t Distribution

Let $x_1, \dots, x_n$ be a random sample from a normal population distribution where σ is unknown. Then the t-statistic

$$t = \frac{\overline{x} - \mu}{SE} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

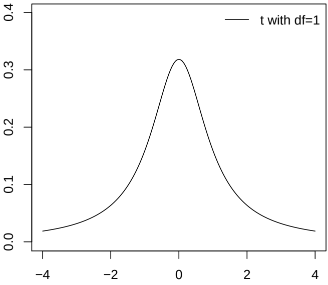
has probability distribution called a t distribution with $n - 1$ degrees of freedom (df).

t Distribution

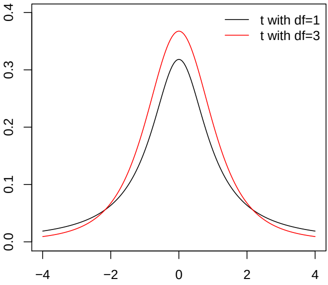
Properties of the t distribution:

- a t-distribution has only one parameter: the degrees of freedom df.
- It is bell-shaped and centered at 0
- Any t curve is more spread out than a z curve.
  i.e. it has fatter tails
- As the df goes to ∞, the t curve approaches the z curve.

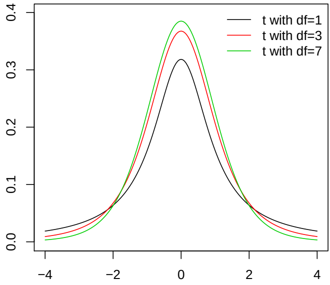
t Distribution Examples



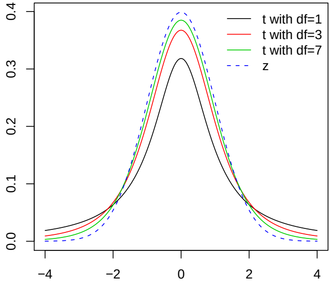
t Distribution Examples



t Distribution Examples



t Distribution Examples



Conditions for Using t Distribution

We use the t distribution when you have

- n is small. E.g. 3, 5, 10, 15.
- Independence: $n \leq 10\%$ rule
- Observations come from a nearly normal distribution:
  - Look at a histogram of the data (difficult when n is small)
  - Consider whether any previous experiences alert us that the data may be normal

t-Tables

If $n = 11$, we use $df = 11 - 1 = 10$ and do a look up on the t-table on page 410:

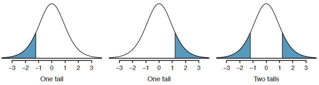


Figure B.1: Three t distributions.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11

Confidence Intervals

Confidence intervals: Use t^*_{df} instead of z^*

$$\overline{x} \pm t^*_{df} SE = \left[\overline{x} - t^*_{df} \times \frac{s}{\sqrt{n}},\ \overline{x} + t^*_{df} \times \frac{s}{\sqrt{n}}\right]$$

So for example, to get a 95% C.I. based on

$$n = 11 \Rightarrow df = 10 \Rightarrow t^*_{10} = 2.23$$

t-Test Example

Example 5.19 on page 252: A random sample of 25 New Yorkers were asked how much sleep they get per night. Does the data below provide strong evidence that New Yorkers sleep less than 8 hours a night on average? Set $\alpha = 0.10$

n	$\overline{x}$	s	min	max
25	7.73	0.77	6.17	9.78

t-Test Example

Conditions:

- Independence: 25 is obviously less than 10% of the population of NYC
- Normality: Not an exact science. The halfway point of the min and max is 7.975, which is fairly close to $\overline{x} = 7.73$. So symmetric enough?

The test statistic is the t-statistic:

$$t = \frac{\overline{x} - \text{null value}}{SE} = \frac{\overline{x} - \text{null value}}{\frac{s}{\sqrt{n}}} = \frac{7.73 - 8}{\frac{0.77}{\sqrt{25}}} = -1.75$$

Since $n = 25$, $df = 25 - 1 = 24$.

t-Test Example

p-Value: we use the t distribution i.e. the t-table on page 410:

one-tail	0.100	0.050	0.025	0.010	0.005
two-tail	0.200	0.100	0.050	0.020	0.010
df = 24	1.32	1.71	2.06	2.49	2.80

Since

- 1.75 is in $[1.71, 2.06]$
- by symmetry -1.75 is in $[-2.06, -1.71]$
- the one-sided p-value is in between $[0.025, 0.05]$

Decision: Since the p-value $< \alpha = 0.10$, we reject H_0 that NY'ers sleep 8 hours a night at the $\alpha = 0.10$ significance level in favor of the hypothesis they sleep more.

History of t Distribution

The t distribution was derived by William Sealy Gosset in 1908, a chemist/statistician at the Guinness Brewery in Dublin, Ireland.

History of t Distribution

Gosset was concerned with small-sample statistics about barley given that brewers are limited in the number of batches of beer they can brew.

Guinness prohibited its employees from publishing. So Gosset had to use the pseudonym "Student" to conceal his identity.

The t-test's complete name is the (Student's) t-test.

History of t Distribution

In fact if you go to the Guinness Brewery at St James's Gate in Dublin, Ireland...

History of t Distribution