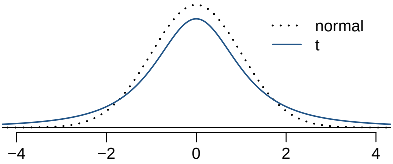
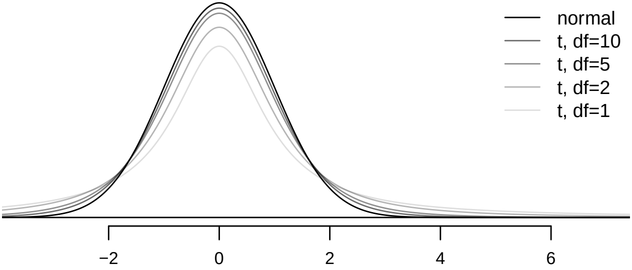
The t distribution

- When working with small samples, and the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.
- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)



The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (df).



What happens to shape of the t distribution as df increases?

Approaches normal.

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a t distribution with $SE = \frac{s}{\sqrt{n}}$.
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - $n < 30$ and no extreme skew
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$

Test statistic

Test statistic for inference on the difference of two small sample means

The test statistic for inference on the difference of two small sample means ($n_1 < 30$ and/or $n_2 < 30$) mean is the T statistic.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

where

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{and} \qquad df = min(n_1 - 1, n_2 - 1)$$

Note: The calculation of the df is actually much more complicated. For simplicity we'll use the above formula to estimate the true df when conducting the analysis by hand.

Recap: Inference using difference of two small sample means

- If $n_1 < 30$ and/or $n_2 < 30$, difference between the sample means follow a t distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, $n <$ 10% of population)
  - independence between groups
  - $n_1 < 30$ and/or $n_2 < 30$ and no extreme skew in either group
- Hypothesis testing:

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = min(n_1 - 1, n_2 - 1)$$

- Confidence interval:

$$\text{point estimate} \pm t^{\star}_{df} \times SE$$