Regression and Other Stories: Ch 4.1 - 4.3 Statistical Inference

"Statistical inference can be formulated as a set of operations on data that yield estimates and uncertainty statements about predictions and parameters of some underlying process or population."

Statistical inference is the process of learning from incomplete data. ROS presents three paradigms for statistical inference.

In many research settings, all three paradigms can apply for a specific situation.

The textbook sets up regression models using the measurement error model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i terms are interpreted as model error (imperfect model specification). Furthermore, the ϵ_i terms are assumed to be a random sample from a probability distribution.

Sampling Distributions / Generative Models The **sampling distribution**

Point Estimates and Uncertainty

This section is, unavoidably, largely a set of definitions.

- **parameters:**

- **estimand:**

- **estimates:**

- **standard error:**

Q: how do the standard error and standard deviation differ?

- **confidence interval:**

- A 95% confidence interval, based on normality assumptions, is roughly $/pm$ 2 standard errors from the mean.

- The standard error for the mean of an infinite population depends on the sample size. Taking a sample of size n from a population with standard deviation σ results in a standard error of σ/n.

Activity:

1. Take a sample of size 100 from a standard normal distribution (`using rnorm()`) and calculate the standard error of the estimate using the standard deviation of the data. How does the result compare to your expected result.

The standard error can be interpreted as the uncertainty in the estimate. One way to think about this uncertainty is the plausible set of estimates from repeatedly running the data generation process.

2. Now, repeat the process and take 1000 replications, each with sample size 100 from a standard normal distribution. Save the mean for each of the 1000 replications and plot this result. Finally take the standard deviation of the sample means, how does this result compare to what you found in part 1?

Now assume that $y \sim N(\mu, \sigma^2)$, then the sampling distribution of the sample mean $\bar{y} = \sum_i y_i/n$. It can be shown that $\bar{y} \sim N(\mu, \sigma^2/n)$.

Similarly, the sample standard deviation, $s_y = \sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2}$, also follows a known distribution,

- **Degrees of freedom:**

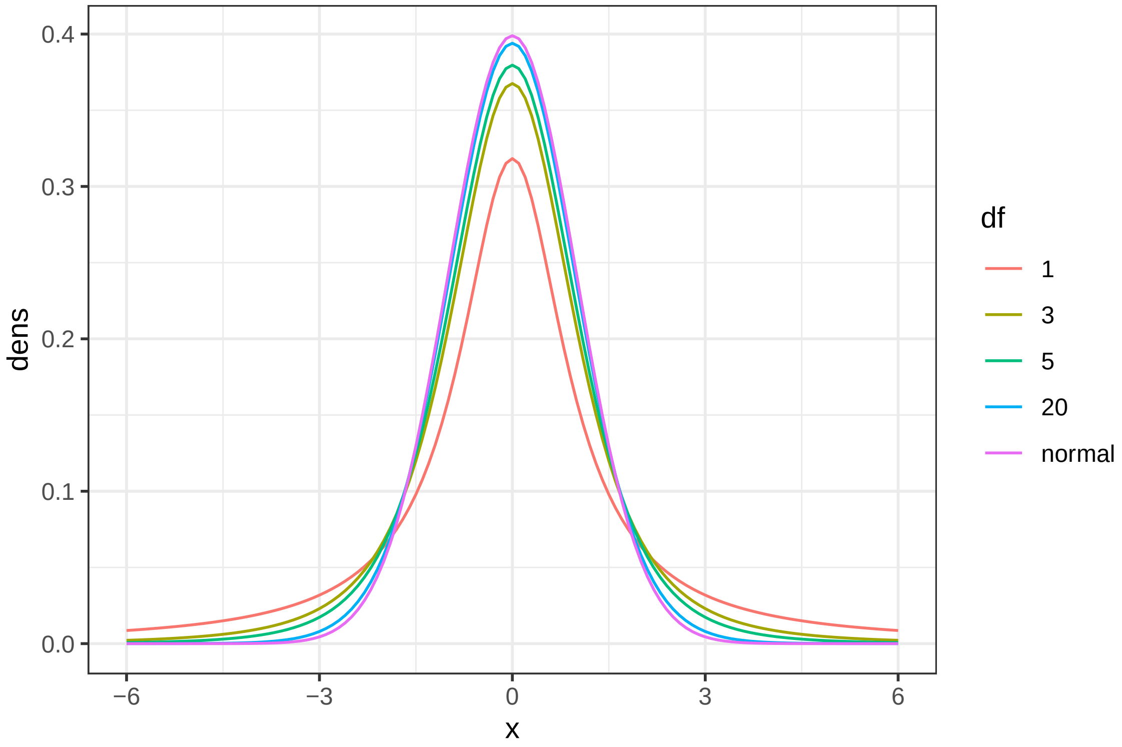
t distributions

- the t-distribution is a set of symmetric distributions characterized by a degree of freedom parameter

- when the degrees of freedom parameter gets large,

- the t-distribution has

- In some cases, the t-distribution is used, along with the standard error, to construct a confidence interval.



Bias (and unmodeled uncertainty)

The preceding inferences require the model being true, having unbiased measurements, and either random samples or randomized experiments.

- An estimate is **unbiased**

Q draw the sampling distribution of an unbiased estimator and a biased estimator.

Unmodeled Uncertainty The textbook presents an example of uncertainty in political polling. Based on the standard error of the binary responses alone, the uncertainty can be very low (likely too low to represent reality).

- How can sources of error not in the model be accounted for?

Inferences will always depend on uncertainty not included in models. Try to reduce some sources and clearly acknowledge assumptions.