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.

1. **Sampling Model:** *goal is to learn characteristics of a population based on a sample. Uncertainty from the sampling process. No reference to measurements.*

2. **Measurement Error Model:** *goal is to learn pattern (functional relationship), but data are measured with error. For instance,* $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. *Can apply even if entire population is surveyed.*

3. **Model error:** *focused on shortcomings of applying imperfect models to observed data.*

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** *is the set of possible datasets that could have been generated if the data collection process had been redone. The uncertainty could arise from any of the three paradigms for statistical inference. A better definition might be the* ***generative model*** *as it represents the random process to generate datasets.*

Point Estimates and Uncertainty

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

- **parameters:** *unknown numbers in a statistical model that is intended to mimic*

- **estimand:** *quantity of interest, could be parameter coefficients or some predicted value*

- **estimates:** *point estimates are best guesses based on data*

- **standard error:** *the estimated standard deviation of an estimate that describes the uncertainty around the estimand.*

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

The standard deviation describes the variability in the data. More data does not change the standard deviation. The standard error describes the uncertainty in an estimate. Collecting more data, results in a more precise estimate.

- **confidence interval:** *a non-precise answer is that a confidence interval represents a range of values for a parameter that are roughly consistent with the data. If the model is correct, then in repeated applications the x% confidence interval will include the true value x% of the time*

- 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.

```
set.seed(09092020)
n <- 100
samples <- rnorm(n)
sd_data <- sd(samples)
se_mean <- sd_data / sqrt(n)
```

The expected result is $\frac{1}{\sqrt{(100)}} = \frac{1}{10}$. The estimated standard error is 0.099.

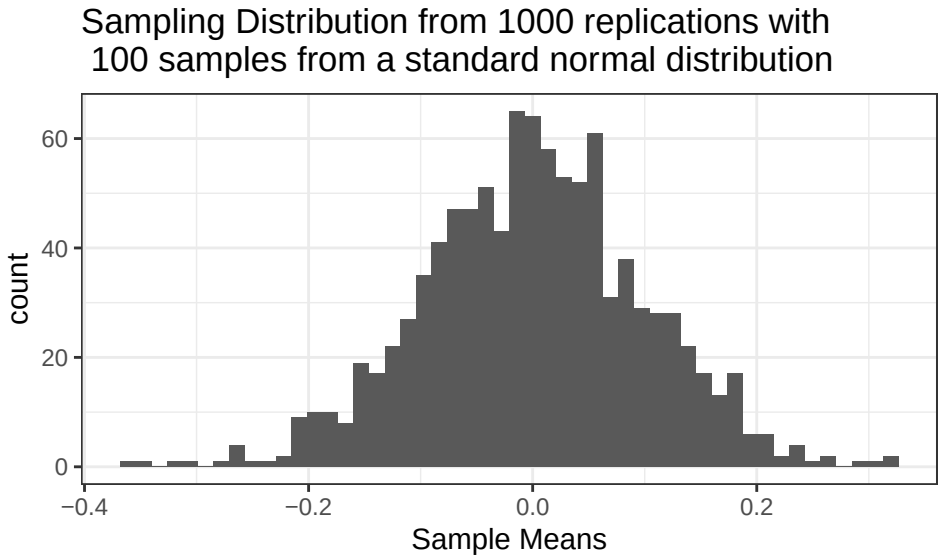
The standard error can be interpreted as the uncertainty in the estimate. One way to think about this uncertainty is the plausible set of estimates from repeatedly running the data generation process.

2. Now, repeat the process and take 1000 replications, each with sample size 100 from a standard normal distribution (`using rnorm()`). 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?

```
library(tidyverse)
replications <- 1000
n <- 100
sample_means <- rep(0, replications)

for (rep in 1:replications){
  sample_means[rep] <- mean(rnorm(n))
}

tibble(sample_means = sample_means) %>%
  ggplot(aes(x = sample_means)) +
  geom_histogram(bins = 50) +
  theme_bw() +
  xlab('Sample Means') +
  ggtitle(paste('Sampling Distribution from', 1000, 'replications with \n', n, 'samples from a standard
```



The standard deviation of the replication means is 0.099 which is very close to the standard error from 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, *the*

$$\chi^2$$

with $n - 1$ *degrees of freedom.*

- **Degrees of freedom:** *the degrees of freedom relate to the number of data points and the number of parameter included in models. When the degree of freedom is low, this adjusts for the uncertainty due to overfitting (or minimal number of data points)*

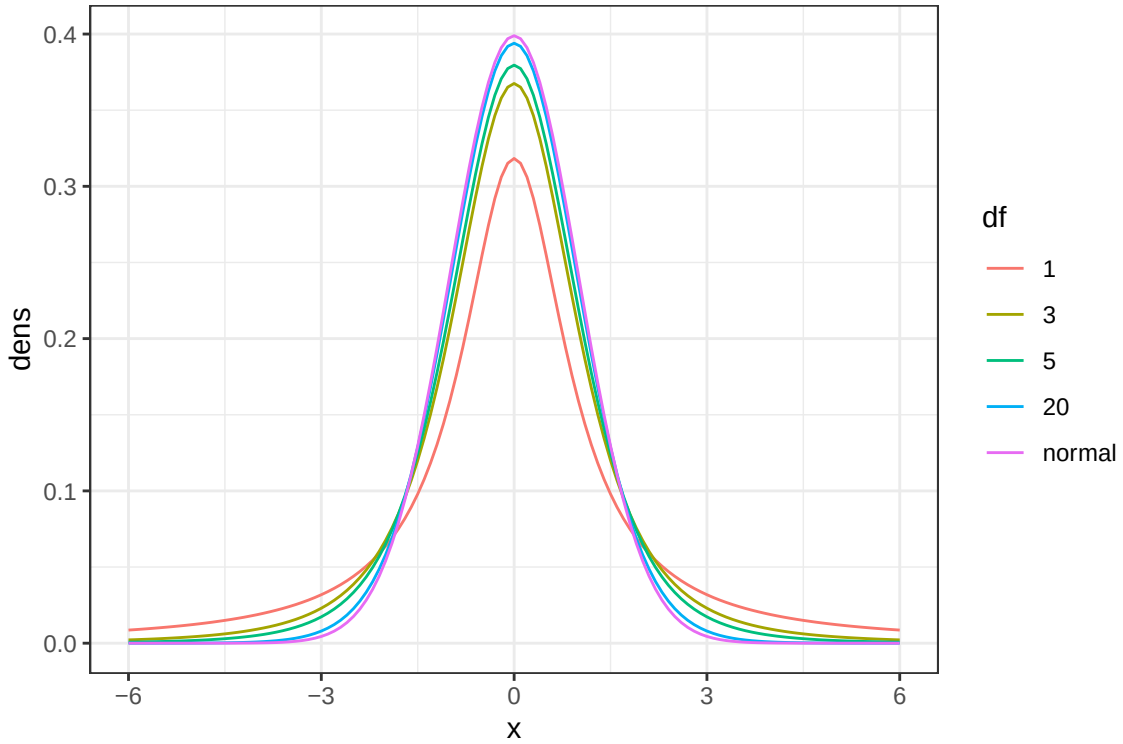
t distributions

- the t-distribution is a set of symmetric distributions characterized by a degree of freedom parameter *(and possible a shift/scale parameter)*

- when the degrees of freedom parameter gets large, *the distribution converges to a normal distribution*

- the t-distribution has *heavier tails than a normal distribution, meaning there is higher probability of extreme values*

- 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. *However, real data collection (measurement) is imperfect, so we need to include model error in inferences.*

- An estimate is **unbiased** *if it is correct on average.*

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?

- *improve data collection: more careful measurements and sampling processes*
- *expand the model: adjust for demographic and geographic information*
- *increase stated uncertainty: known nonsampling error in polling dat of 2.5%*

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