Social Science Inquiry II

Week 6: Linear models, part II

Molly Offer-Westort

Department of Political Science,
University of Chicago

Winter 2023

Loading packages for this class

```
> library(ggplot2)
> library(estimatr)
> set.seed(60637)
```

Returning to inference

- We have some data that are produced from a random sampling procedure, where they are sampled from the same population.
- We've selected an estimating procedure, and produced a point estimate of some target estimand using our estimating procedure.
- We then produced an estimate of the standard error of our estimate.
- Now we would like to be able to say something what that means.

One way to do this is to use our estimated standard errors to give an interval of uncertainty around our point estimate.

Confidence intervals

- A valid confidence interval CI_n for a target parameter θ with coverage $1 - \alpha$

$$\mathrm{P}[\theta \in CI_n] \geq 1 - \alpha$$

- If $\alpha = 0.05$, the probability that the estimand θ is in our confidence interval is greater than or equal to 0.95.
- CI_n is a random interval. It is a function of the data we observe.
- θ is a fixed parameter. It does not move.
  (In the frequentist view of statistics.)
- If you use valid confidence repeatedly in your work, 95% of the time, your confidence intervals will include the true value of the relevant θ.

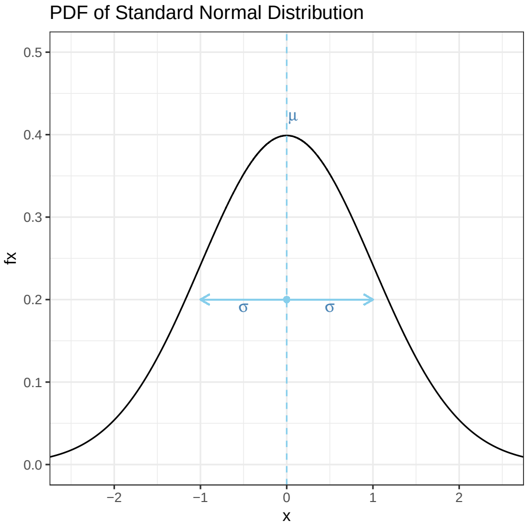
- We could trivially define valid confidence intervals by including the entire support of the data. (Why wouldn't we want to do that?)

- The formula for the 95% confidence interval is:

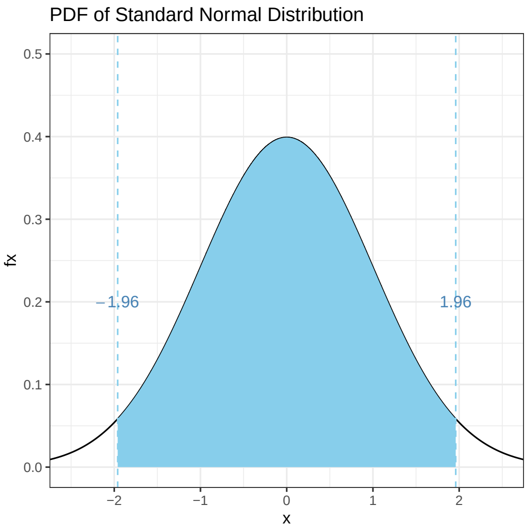
$$CI_n = \left(\hat{\theta}_n - 1.96 \times \hat{\mathrm{se}},\ \hat{\theta}_n + 1.96 \times \hat{\mathrm{se}}\right)$$

- The 1.96 value tells us how many standard errors away from the mean we need to include in our interval to get valid coverage.
- This formula is based on a *normal approximation*, i.e., we assume the data is going to look like a normal distribution.

The normal distribution has a bell curve shape, with more density around the middle, and less density at more extreme values.



If we want to describe symmetric bounds around the mean that contain 95% of the distribution, this would be from the 2.5th percentile to the 97.5th percentile.



Returning to our example where we flip a coin twice, let X be the number of heads we observe. Our coin is *not* fair, and the probability of getting a heads is 0.75.

```
> X <- c(0, 1, 2)
> fx <- c(1/16, 3/8, 9/16)
> (Ex <- sum(X*fx))
```

```
[1] 1.5
```

Let's take a sample of size $n = 100$ from this distribution, and see what our confidence intervals look like.

```
> n <- 100
> x_observed <- sample(X, prob = fx, replace = TRUE, size = n)
> head(x_observed)
```

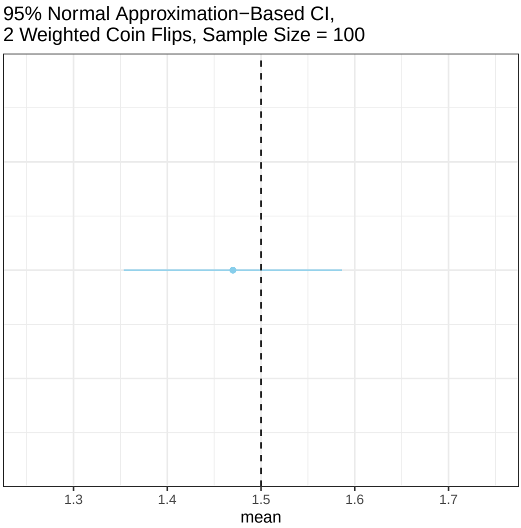
```
[1] 1 2 1 1 1 2
```

Our estimates of the mean and standard error of the mean.

```
> (theta_hat <- mean(x_observed))
[1] 1.47
> (se_hat <- sd(x_observed)/sqrt(length(x_observed)))
[1] 0.05938234
```

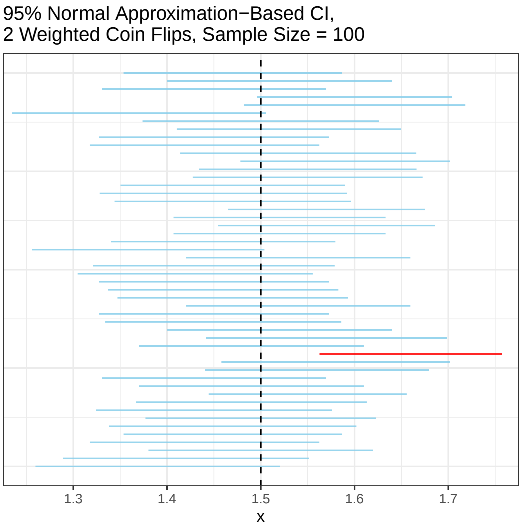
Putting it together, the 95% normal approximation-based confidence interval.

```
> (CI95 <- c(theta_hat + c(-1,1)*1.96*se_hat))
[1] 1.353611 1.586389
```


95% Normal Approximation−Based CI,
2 Weighted Coin Flips, Sample Size = 100
1.3
1.4
1.5
1.6
1.7
mean

What if we did this many times?

```
> n_iter <- 50
> x_mat <- replicate(n_iter, sample(X, prob = fx, replace = TRUE,
+                                   size = n))
> CI_95f <- function(x){
+   theta_hat <- mean(x)
+   se_hat <- sd(x)/sqrt(length(x_observed))
+   CI_hat <- theta_hat +
+     c('conf_lower' = -1, 'conf_upper' = 1)*1.96*se_hat
+ }
> sample_CIs <- as.data.frame(t(apply(x_mat, 2, CI_95f)))
> head(sample_CIs, 3)
  conf_lower conf_upper
1   1.259646   1.520354
2   1.288800   1.551200
3   1.380177   1.619823
```


95% Normal Approximation−Based CI,
2 Weighted Coin Flips, Sample Size = 100
1.3
1.4
1.5
1.6
1.7
x

- The true mean stays the same.
- The confidence intervals change, based on the sample.

```
> mean( (Ex >= sample_CIs$conf_lower) &
+         (Ex <= sample_CIs$conf_upper) )

[1] 0.98
```

What if we did this many more times?

```
> x_mat <- replicate(5000, sample(X,
+                                  prob = fx,
+                                  replace = TRUE,
+                                  size = n))
> CI_n <- as.data.frame(t(apply(x_mat, 2, CI_95f)))
> mean( (Ex >= CI_n$conf_lower) & (Ex <= CI_n$conf_upper) )
[1] 0.9508
```

Applied example

We can see this in action with respect to the paper by Devah Pager:

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology*, 108(5), 937-975.

- The study was an audit study, where pairs of white and pairs of black hypothetical job applicants applied to real jobs.
- In each pair, one respondent listed a criminal record on job applications; the other did not. Otherwise, applicants were matched.
- The outcome is whether applicants got a callback.

```
> dfp <- data.frame(
+   black = rep(c(0, 1), times = c(300, 400)),
+   record = c(rep(c(0, 1), each = 150),
+              rep(c(0, 1), each = 200)),
+   call_back = c(
+     # whites without criminal records
+     rep(c(0, 1), times = c(99, 51)), # 150
+     # whites with criminal records
+     rep(c(0, 1), times = c(125, 25)), # 150;
+     # - callbacks could be 25 or 26
+     # blacks without criminal records
+     rep(c(0, 1), times = c(172, 28)), # 200
+     # blacks with criminal records
+     rep(c(0, 1), times = c(190, 10)) # 200
+   )
+ )
>
```

```r
> pager_agg <- aggregate(call_back~black + record, data = dfp, mean)
> pager_agg$race <- factor(pager_agg$black,
+                          levels = c(1, 0),
+                          labels = c('Black', 'White'))
> pager_agg$criminal_record <- factor(pager_agg$record,
+                                     levels = c(1, 0),
+                                     labels = c('Record', 'No Record'))
> pager_agg
  black record call_back  race criminal_record
1     0      0 0.3400000 White       No Record
2     1      0 0.1400000 Black       No Record
3     0      1 0.1666667 White          Record
4     1      1 0.0500000 Black          Record
```

American Journal of Sociology

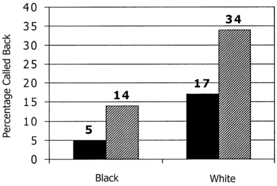


FIG. 6.—The effect of a criminal record for black and white job applicants. The main effects of race and criminal record are statically significant ($P < .01$). The interaction between the two is not significant in the full sample. Black bars represent criminal record; striped bars represent no criminal record.

Let's say our $\hat{\theta}$ here is the overall mean of `call_back` among black applicants.

```
> (theta_hat <- mean(dfp$call_back[which(dfp$black == 1)]) )
[1] 0.095
```

Our $\hat{se}$ is our estimate of the standard error of the mean,

$$\hat{se} = \sqrt{\hat{\mathrm{Var}}[X]/n}.$$

We get this by plugging in our unbiased sample variance estimate into the formula for the standard error of the mean.

```
> (se_hat <- sqrt(var(dfp$call_back[which(dfp$black == 1)])/
+                     length(which(dfp$black == 1))))
[1] 0.01467911
>
>
```

We can then get our 95% confidence intervals by plugging into the formula,

$$CI_n = \left(\hat{\theta}_n - 1.96 \times \hat{\mathrm{se}},\ \theta_n + 1.96 \times \hat{\mathrm{se}}\right)$$

```
> (CI <- c(theta_hat + c(-1,1)*1.96*se_hat))
[1] 0.06622895 0.12377105
```

Inference for linear models

- As a special case of a linear model, when we regress Y on an indicator, we just get the sample mean of Y.
- In this case, estimating standard errors and confidence intervals follows the same procedures as for sample means.

```
> model <- lm_robust(call_back ~1,
+             data = dfp[which(dfp$black == 1),])
```

Inference for linear models

```
> summary(model)

Call:
lm_robust(formula = call_back ~ 1, data = dfp[which(dfp$black ==
    1), ])

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper
(Intercept)    0.095    0.01468   6.472 2.844e-10  0.06614   0.1239

Multiple R-squared:  5.218e-15 ,        Adjusted R-squared:  5.218e
```

Inference for linear models

```
> confint.default(model)

                 2.5 %    97.5 %
(Intercept) 0.06622948 0.1237705
```

- We can think about parameters in a linear model in a similar way.

$$\mathrm{E}[Y|X] = \beta_0 + \beta_1 X$$

- The true population parameters are generally unknown.

- We estimate them for a given sample.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- We will think about our random sample being not just for one variable, but from the joint distribution of (Y, X).
- Then each $\hat{\beta}_k$ is also random, with its own sampling distribution.
- We can get a point estimate for each of the parameters, $\hat{\beta}_k$: the coefficients in our linear model.
- We also want to get an estimate of the standard errors of the estimates, $\sqrt{\hat{\mathrm{Var}}[\hat{\beta}_k]}$.

- Let's try this with the dfp data, where the outcome Y is `call_back`, regressed on `black`.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Black}_i$$

```
> model2 <- lm_robust(call_back ~ black, data = dfp)
```

- How do we go about interpreting these coefficients and confidence intervals?

```
> summary(model2)

Call:
lm_robust(formula = call_back ~ black, data = dfp)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
(Intercept)   0.2533    0.02515  10.072 2.270e-22   0.2040   0.3027 698
black        -0.1583    0.02912  -5.437 7.504e-08  -0.2155  -0.1012 698

Multiple R-squared:  0.04503 ,        Adjusted R-squared:  0.04366
F-statistic: 29.56 on 1 and 698 DF,  p-value: 7.504e-08
```

- How do we go about interpreting these coefficients and confidence intervals?

```
> confint.default(model2)
                 2.5 %     97.5 %
(Intercept)  0.2040362  0.3026305
black       -0.2154118 -0.1012548
```

References I

Pager, D. (2003). The mark of a criminal record. American journal of sociology, 108(5):937–975.