Quantitative Methods – Day 1: Quantifying Uncertainty

Harm H. Schuett

08 January, 2018

Contents


```
# imports
library(ggplot2)
library(gridExtra)
```

Recap: What is uncertain about our estimates?

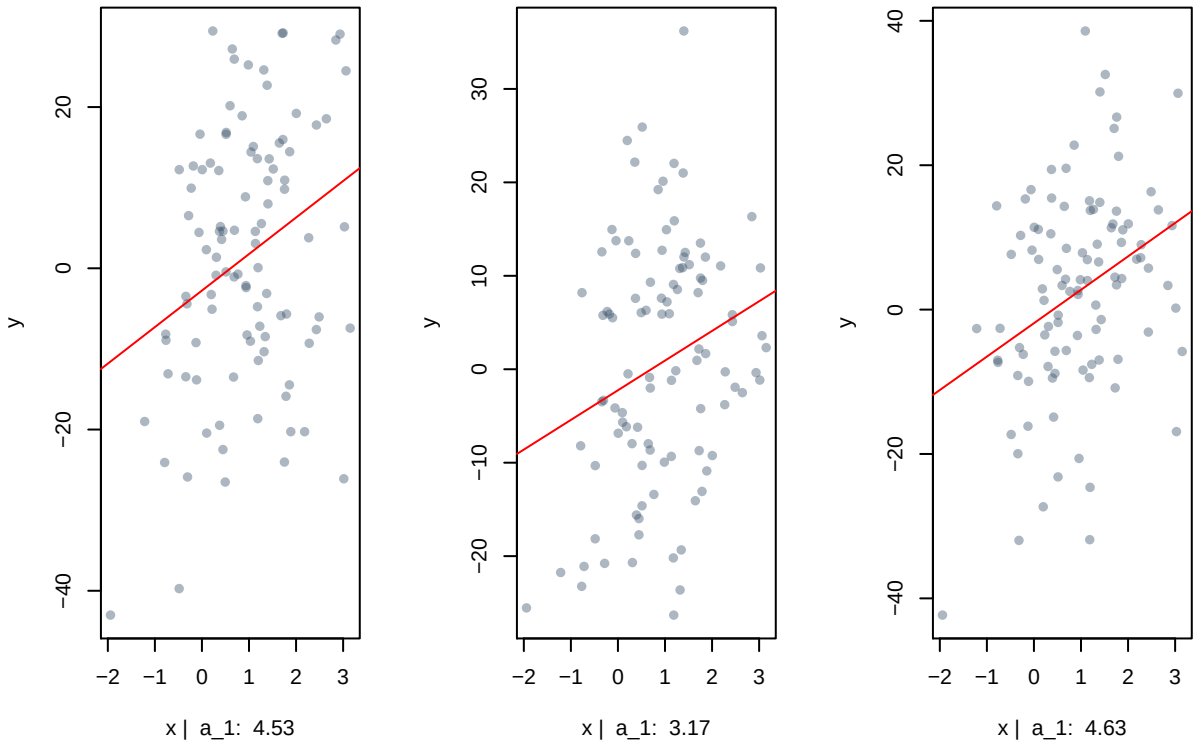
Let's quickly recapitulate what we discussed in the last notebook. If we would replicate the same data but have unmeasured influences u, then we will get different estimates. Assume again:

$$y_i = a_0 + a_1 * x_i + u_i$$

where $u_i = 3 * z_i - 2 * v_i$. Let's look at three tests:

```
set.seed(666)
n <- 100
reps <- 3
x <- rnorm(n, mean=1, sd=1)  # don't change x
# a1_sim <- vector(mode="numeric", length=reps)
par(mfrow=c(1, 3))  # 3 plots side-by-side
for (i in 1:reps){  # redraw unmeasured z, and v:
  z <- rnorm(n=n, mean=1, sd=4)
  v <- rnorm(n=n, mean=2, sd=4)
  y = 1 + 2 * x + 3 * z - 2 * v
  fit <- lm(y~x)
  a1 <- coef(fit)["x"]
  plot(x=x, y=y,
       col=rgb(0.2, 0.3, 0.4, 0.4), pch=16,
       xlab=paste("x |  a_1: ", round(a1,2)))
```

```
  abline(fit, col="red")
}
```



The variation is due to the unmeasured influence of z and v. The world is too complex for us to track all influences for nearly all variables of interest and we will never be able to measure all influences. When we try to draw inferences from just one sample, we need to figure out how uncertain our estimates are. Depending if you think in Bayesian or Frequentist terms, this uncertainty is either expressed directly as uncertainty due to missing information or in terms of hypothetical variation from imagined repeat of the experiment (which is how we framed our analysis so far). Both approaches are great, have advantages and disadvantages. Fundamentally, they just are different ways to handle the same problem: quantifying uncertainty. Let's talk about the classic frequentist approach, since this is still the de facto standard.

Bias, variance and the error term

Let's look at the OLS estimator closely again:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

We can expand this and take a closer look why changes in u affect the estimate $\hat{\beta}$ of β.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1}$$
$$= (X'X)^{-1}X'(X\beta + u) \tag{2}$$
$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \tag{3}$$
$$= \beta + (X'X)^{-1}X'u \tag{4}$$
$$\tag{5}$$

Quite obviously, the estimate $\hat{\beta}$ is equal to the actual β (if the relationship is indeed linear) plus the additional $(X'X)^{-1}u$ term. That means that in reality, we will never measure the true β. But this tells us two things:

If we want to be as close as possible to the true β we would like to:

1. Hit the true β at least on average if we could do repeated experiments
2. Have the term $(X'X)^{-1}u$ as small as possible so that we do not stray far away from the true β

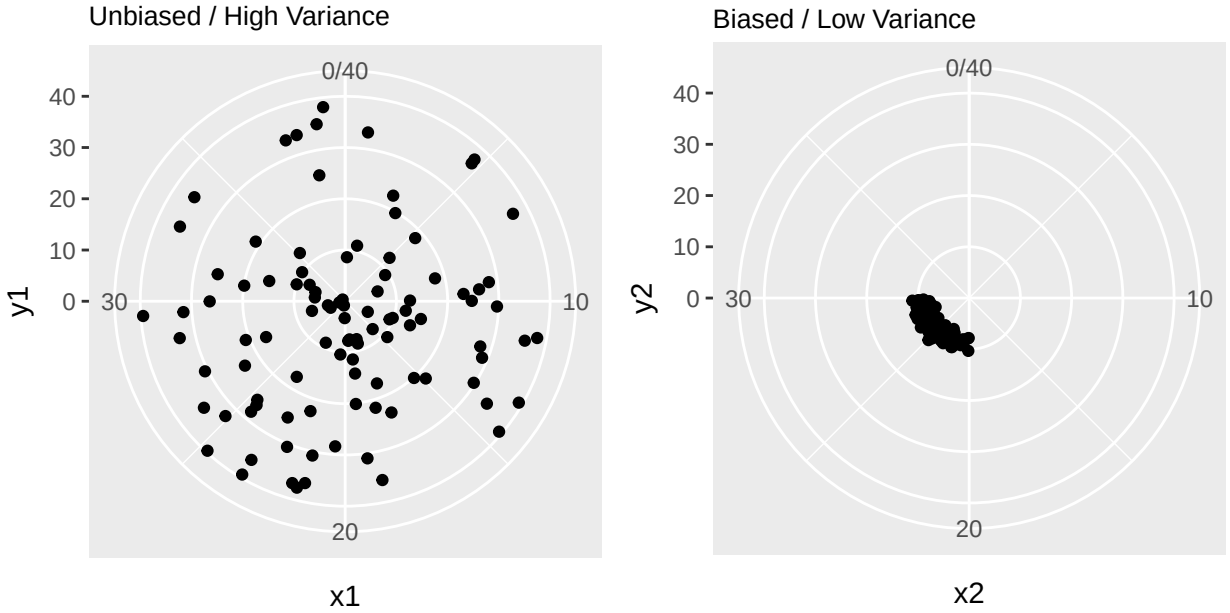
Regarding point one:

$$E[\hat{\beta}] = E[\beta + (X'X)^{-1}X'u]$$

If $E[X'u] = 0$, then $E[\hat{\beta}] = \beta$. This is sometimes called **exogeneity assumption**. We need this to hit β at least on average. Stated another way, we need this assumption to have an *unbiased* estimator. We will talk a lot about bias in following sessions.

The second desirable property of an estimator is low *variance*. Both *unbiasedness* and *low variance* are important. There is an incredibly nerdy poem called "Hiawatha designs an experiment", which is about a native american who learns to shoot unbiased arrows from the best statisticians in the world. Problem is that even though he would hit the target on average he sprays widely, never hits the target, and his tribe takes away his bow and arrows.

```
df <- data.frame(x1 = runif(100, min=0, max=40),
                 x2 = runif(100, min=20, max=30),
                 y1 = runif(100, min=0, max=40),
                 y2 = abs(rnorm(100, mean=9, sd=1)))
# theme_set(theme_bw())
left_plot <- ggplot(data=df, aes(x1, y1)) +
  geom_point() +
  coord_polar() +
  scale_y_continuous(limits=c(0, 40)) +
  scale_x_continuous(limits=c(0, 40))+
  labs(subtitle="Unbiased / High Variance")
right_plot <- ggplot(data=df, aes(x2, y2)) +
  geom_point() +
  coord_polar() +
  scale_y_continuous(limits=c(0, 40)) +
  scale_x_continuous(limits=c(0, 40)) +
  labs(subtitle="Biased / Low Variance")
grid.arrange(left_plot, right_plot, ncol=2)
```



Looking at the two different dart boards, which bowman would you prefer? In econometrics, everyone searches for a minimum-variance unbiased estimator (MVUE).

One can show that if the u_i behave in a certain way ($\sigma^2(u_i) = \sigma^2 \ \forall i$ and $\sigma(u_i, u_j) = 0 \ \forall i \neq j$) then OLS is indeed the minimum variance estimator. But to see this, we need to find a way to quantify the variance of an estimator.

Coefficient standard errors

Going back to $\hat{\beta} = \beta + (X'X)^{-1}X'u$, we can try to quantify this estimator's variance:

$$
\begin{aligned}
var[\hat{\beta}|X] &= E[((X'X)^{-1}X'u)((X'X)^{-1}X'u)'|X] && (6)\\
&= E[(X'X)^{-1}X'uu'X(X'X)^{-1}|X] && (7)\\
&= \underbrace{(X'X)^{-1}X'}_{\text{Weights}} \quad \underbrace{E[uu'|X]}_{\text{Variance-Covariance of u}} \quad \underbrace{X(X'X)^{-1}}_{\text{Weights}} && (8)\\
& && (9)
\end{aligned}
$$

So, here again you can see, the variance of $\hat{\beta}$ (figuratively speaking: the amount of dart board that the estimator is spraying) is a function of the $E[uu'|X]$, variance-covariance matrix of the error term u. The more important the unmeasured influences in u in terms of variation contribution for Y, the more it will interfere with measuring the relation between Y and X. The joint effects of the $(X'X)^{-1}X'$ parts on both sides are similar in spirit to the scalar expression $\cdot \frac{x}{x^2}$, which means the less variation there is in X, the less

precise is $\hat{\beta}$. However, what is also important in $(X'X)^{-1}X'$ is the covariance between the different variables – x_1, x_2, x_3, etc. – comprising X. This is where *colinearity* comes in, but we will talk about that later.

The key term remains $E[uu'|X]$, which given the focus of the whole discussion on u – the unmeasured influences – so far shouldn't surprise you anymore. This is a matrix with has the form:

$$E[uu'|X] = \begin{pmatrix} \sigma_{u_1}^2 & \sigma_{u_1,u_2} & \dots & \sigma_{u_1,u_n} \\ \sigma_{u_1,u_2} & \sigma_{u_2}^2 & \dots & \sigma_{u_2,u_n} \\ \dots & \dots & \dots & \dots \\ \sigma_{u_1,u_n} & \sigma_{u_2,u_n} & \dots & \sigma_{u_n}^2 \end{pmatrix} \tag{10}$$

This matrix is what really drives the shape of the (co-)variance of the coefficients in $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots)$. The problem is of course that we do not really know how this matrix looks like. **We can only make assumptions about its shape**. We will spend a whole session talking about what to assume here, depending on the situation (aka clustered standard errors and the like). But for now, we will continue with the classic text book example of **homoscedasticity**. Assume:

$$E[uu'|X] = \begin{pmatrix} \sigma_u^2 & 0 & \dots & 0 \\ 0 & \sigma_u^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_u^2 \end{pmatrix} = \sigma_u^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma_u^2 I \tag{11}$$

I words, this says that the unmeasured influences spray the same amount of noise onto the relation between Y and X irrespective of what observation we are looking at and the noise on one observation does not affect the noise in another. Plugging this into the $var[\hat{\beta}|X]$ yields:

$$var[\hat{\beta}|X] = (X'X)^{-1}\sigma_u^2 I X(X'X)^{-1} = \sigma_u^2 (X'X)^{-1}$$

Hypothesis testing and p-values

Classic hypothesis testing as a decision tool

Now we have a measure of how precise or imprecise the estimate is. But we still we need to figure out how to use it. One way is to use classic hypothesis testing, which is basically a decision theoretic approach to judge whether an estimator is too imprecise or not to learn something from it.

Set up a *null hypothesis* and an *alternative hypothesis*. For instance, often the null is $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$. What we want to know is whether we can reject H_0, i.e., whether it is very unlikely, given the precision of our estimates, that the data our estimator saw could have been produced by a $\beta = 0$. Hypothesis testing via p-values is a way of trying to give you a decision theoretically sound formula for such a decision, But you need to specify the amount of possible error you are willing to incur if you make a decision. Also, p-values have been criticized very heavily in recent years. We will discuss how p-values work and what the virtues and pitfalls are next.

From standard errors to hypothesis testing

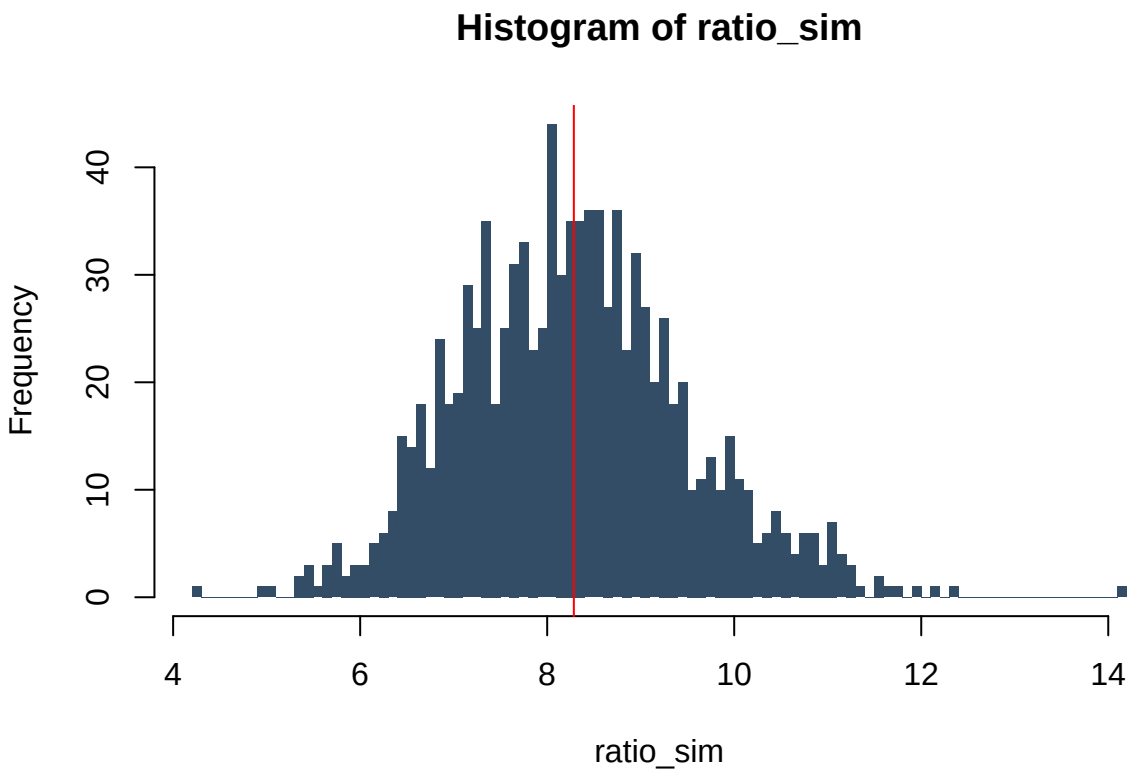
To repeat again what we are after: we want to know whether we can safely reject H_0. That is, whether it is very unlikely, given the precision of our estimates, that the data our estimator saw could have been produced by a $\beta = 0$. All the information we have to answer this question is our estimate $\hat{\beta}$ and its variance $\sigma(\hat{\beta})$. ...

Or do we? Actually, we do not have this variance. That one is unknown to us too, since we cannot really do this hypothetical experiment of repeating our tests and we do not know how $E[uu'|X]$ really looks like. So we need to estimate this variance from the data too and using some assumption about $E[uu'|X]$, like the one we did above: $\hat{\sigma}(\hat{\beta})$.

Now, armed with $\hat{\beta}$ and $\hat{\sigma}(\hat{\beta})$, how can we figure out how likely $\beta = 0$ is? We need to take into account the precision of our estimator. One way of doing this is to construct a ratio $\hat{\beta}/\hat{\sigma}(\hat{\beta})$. Lower values will either be due to small magnitude estimates ($\hat{\beta}$) or imprecise estimation ($\hat{\sigma}(\hat{\beta})$). Similarly, a very high magnitude of the estimate, even if imprecisely estimated would still give a comparably high value of this ratio. We can thus make an argument that if this ratio is sufficiently far away from zero, then we can reject the possibility that $\beta = 0$; at least in theory. Of course, we need to operationalize what we mean by "sufficiently" and here is where p-values come into play.

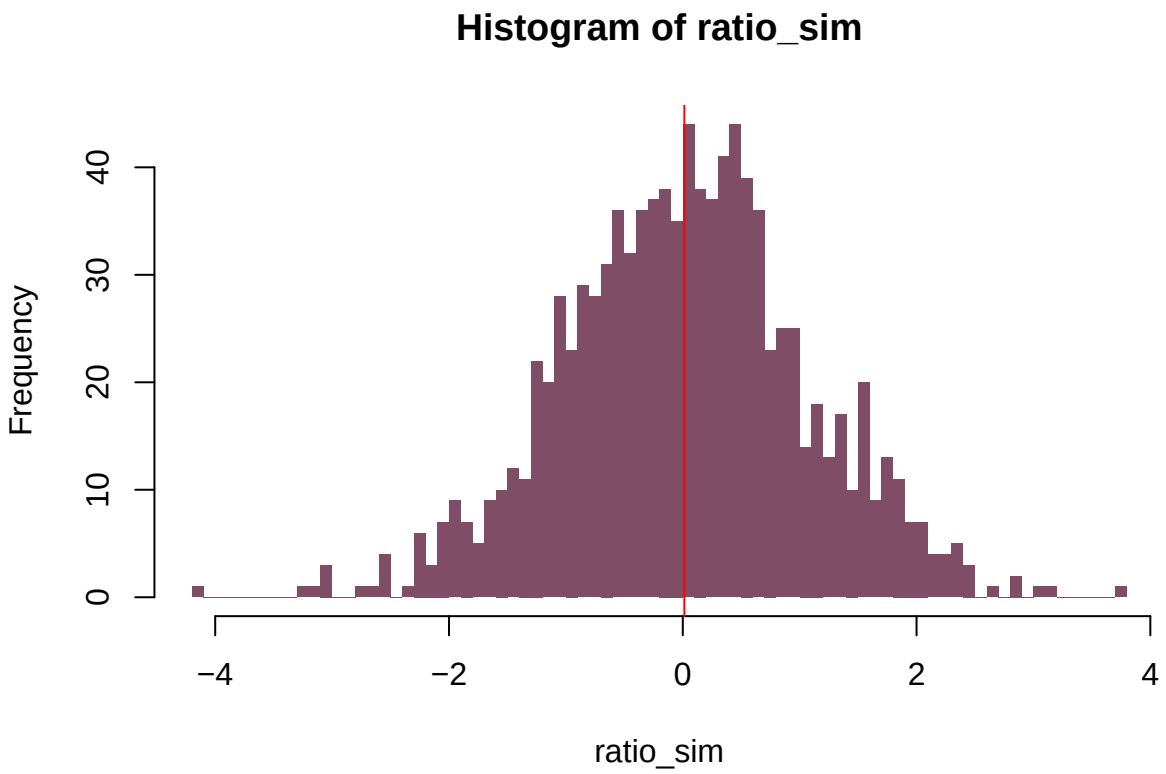
Remember that if you had 1000 hypothetical replications of your data, you would get 1000 different $\hat{\beta}$ because of the unmeasured influences u. The same goes for your estimate $\hat{\sigma}(\hat{\beta})$ since you need to estimate those from the data as well. Consequently, you would get 1000 ratios $\hat{\beta}/\hat{\sigma}(\hat{\beta})$. Let's simulate that:

```
set.seed(666)
reps <- 1000
n <- 100
x <- rnorm(n=n, mean=1, sd=2)
ratio_sim <- vector(mode="numeric", length=reps)
for (i in 1:reps){
  y <- 1 + 2 * x + rnorm(n=n, mean=0, sd=5)
  fit <- lm(y ~ x)
  betas <- coef(fit)
  ses <- coef(summary(fit))[, "Std. Error"]
  ratio_sim[i] <- betas["x"] / ses["x"]
}
hist(ratio_sim, col=rgb(0.2, 0.3, 0.4, 1), border=NA, breaks=100)
abline(v=mean(ratio_sim),col="red")
```



So, in this case, the true β is 2 and we can see the ratio is around 8 on average (the estimator is reasonably precise). But we still need a benchmark. Think a bit about the following: What we actually need to know for reasoning about whether $H_0 : \beta = 0$ is *what the above histogram would look like, if* $\beta = 0$.

```
set.seed(666)
reps <- 1000
n <- 100
x <- rnorm(n=n, mean=1, sd=2)
ratio_sim <- vector(mode="numeric", length=reps)
for (i in 1:reps){
  y <- 1 + 0 * x + rnorm(n=n, mean=0, sd=5)
  fit <- lm(y ~ x)
  betas <- coef(fit)
  ses <- coef(summary(fit))[, "Std. Error"]
  ratio_sim[i] <- betas["x"] / ses["x"]
}
hist(ratio_sim, col=rgb(0.5, 0.3, 0.4, 1), border=NA, breaks=100)
abline(v=mean(ratio_sim),col="red")
```



Obviously, the ratios that we would get from our hypothetical data replications would be centered around zero (because OLS is an unbiased estimator and the true beta is zero). What is more interesting is the spread of this curve. It is reasonably symmetric and most of its mass is between -2 and +2. So, a ratio of around 8 (as in the previous simulation with $\beta = 2$) would be very unlikely if the true $\beta = 0$. It is possible, but very very unlikely. So unlikely in fact, that we can reject $\beta = 0$ in favor of $\beta \neq 0$ and be confident that the probability of us falsely rejecting $\beta = 0$ is very small.

But, at which magnitude of this ratio should we start saying that $\beta = 0$ becomes very unlikely? We would like to know the probability of making an error if we decide to reject H_0. Something like, "if my estimated ratio $\hat{\beta}/\hat{\sigma}(\hat{\beta})$ is 8 and I reject $\beta = 0$, then the chance of making an error because β truly is null is p%". Notice that we could do this simply from the red histogram above. Similarly, if we would know what *ditribution* the estimated ratios follow if $\beta = 0$, we could calculate that probability. And with some assumptions we can figure this out. We won't go into the derivations here, you can read them up in any econometric text book. But assuming normally distributed errors, it is easy to show that our estimate $\hat{\beta}$ is normally distributed with:

$$\hat{\beta} \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$$

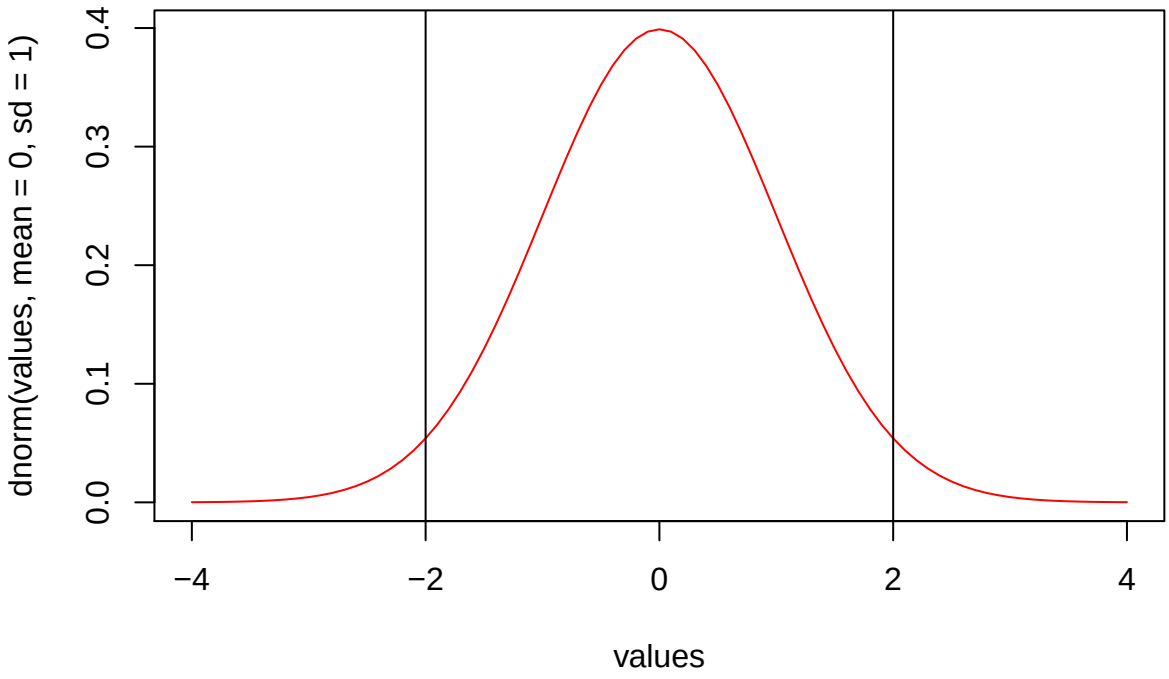
If we knew σ, we would be done already. Per rules for normal distributions, if you subtract the mean and scale a normally distributed random variable by its standard distribution, you get the standard normal distribution:

$$\hat{\beta} \sim N\left(\beta, \sigma^2(X'X)^{-1}\right) \rightarrow \frac{\hat{\beta} - \beta_0}{\sigma(\beta)} \sim N(0, 1)$$

β_0 is the assumed true β under the null. In our discussion we assumed $\beta_0 = 0$, but we can assume other values too. The idea is exactly the same. Once you subtract that value, you end up with a standard normal

distribution. And we know how the standard normal distribution looks like:

```
values <- seq(-4,4,0.1)
plot(x=values, y=dnorm(values, mean=0, sd=1), type="l", col="red")
abline(v=c(-2, 2))
```



With values greater 1.96 *or less* occurring with probability:

```
pnorm(1.96, mean=0, sd=1)
```

```
## [1] 0.9750021
```

So, we know that there is a 2.5% chance of seeing a ratio greater than 1.96 if our ratios where drawn from a standard normal distribution. Unfortunately, they are not since we do not know σ^2 in $\sigma^2(X'X)^{-1}$. We need to estimate it.

A consistent estimator for $\sigma^2(X'X)^{-1}$ is

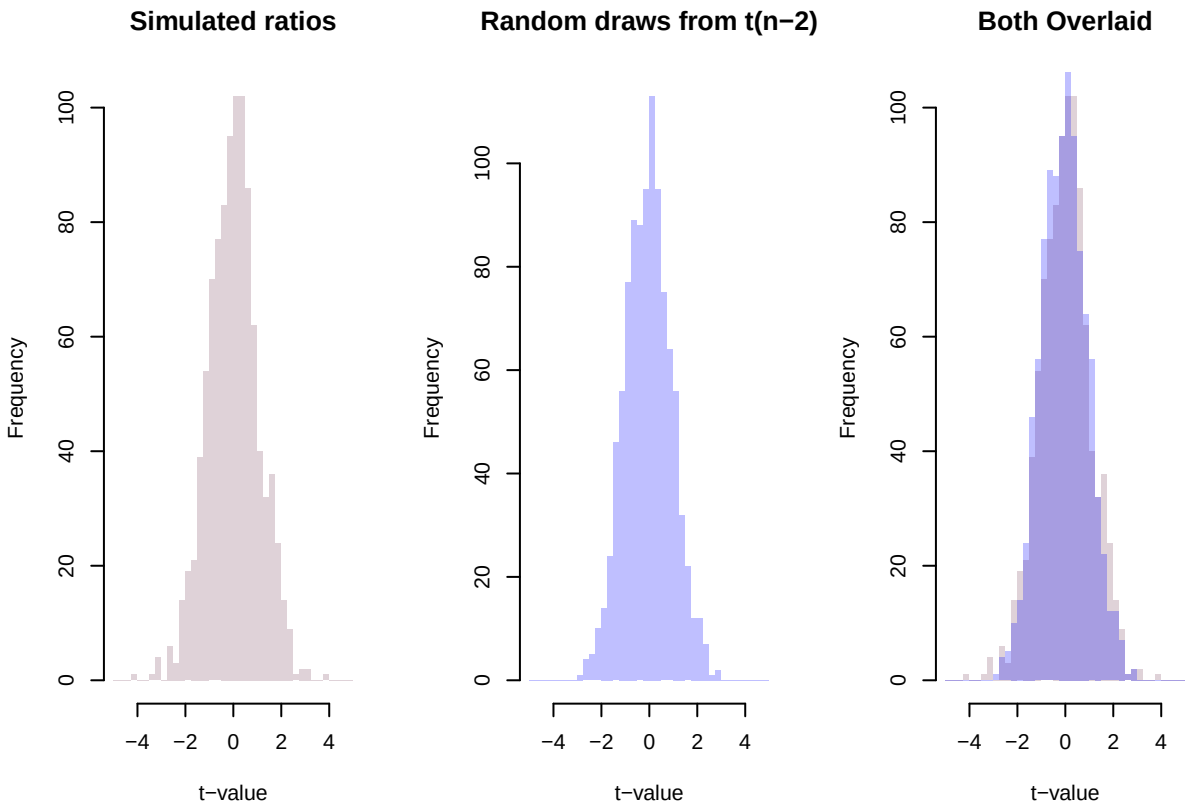
$$\hat{\sigma^2} = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2 \rightarrow se(\hat{\beta}) = \sqrt{\hat{\sigma^2}(X'X)^{-1}}$$

Since it is estimated from data, $\hat{\sigma^2}$ has a distribution by itself, a chi-squared distribution. Thus, the ratio is a ratio of a normally distributed random variable to the square root of an independent chi-2 variable, which gives a t distribution (See f.e., Goldberger (1991) pp. 223):

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \sim t_{n-k}$$

Which, is why this ratio is usually called a t-statistic. Let's see whether we can visualize this to get some more intuition.

```
set.seed(666)
par(mfrow=c(1,3))
breaks <- seq(-5, 5, 0.25)
cred <- rgb(0.5, 0.3, 0.4, 0.25)
cblue <- rgb(0,0,1,1/4)
# 1st histogram
hist(ratio_sim, col=cred, border=NA, breaks=breaks,
     main="Simulated ratios", xlab="t-value")
# drawing randomly from the t-distribution
t_dist <- rt(n=reps, df=n-2)
# 2nd histogram
hist(t_dist, col=cblue, border=NA, breaks=breaks,
     main="Random draws from t(n-2)", xlab="t-value")
# 3rd histogram
hist(ratio_sim, col=cred, border=NA, breaks=breaks,
     main="Both Overlaid", xlab="t-value")
hist(t_dist, col=cblue, border=NA, breaks=breaks, add=T)
```



As you can see, random draws from the t-distribution fits our simulated ratios very well. Going back to the previous argument, if our estimated $t_{\hat{\beta}}$ follows a t-distribution, we can easily use that distribution to say how likely it is to see a value $t_{\hat{\beta}} = 8$ or less if $\beta = 0$

```
# using the t-probability function to compute the prob
# to see a value of 8 or less
```

```r
pt(q=8, df=n-2)
```

```
## [1] 1
```

Well, (due to rounding) the probability of seeing a value of 8 or less is 1. Let's do something more classic: Let's compute the probability of seeing a t-value of 2 or less:

```r
# using the t-probability function to compute the prob
# to see a value of 8 or less
pt(q=2, df=n-2)
```

```
## [1] 0.9758661
```

So that probability is roughly 97.5% and the probability of seeing a value of 2 *or more* is 2.5% respectively. Since the t-distribution is symmetric, the same goes for -2.

Problems when interpreting confidence intervals or p-values

> "... the distance between the data and the model prediction is measured using a test statistic (such as a t-statistic or a Chi squared statistic). The P value is then the probability that the chosen test statistic would have been at least as large as its observed value if every model assumption were correct, including the test hypothesis. ... the P value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility, and in this sense may be viewed as measuring the fit of the model to the data." — Greenland, Sander, et al. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." European Journal of Epidemiology 31 (2016), p. 339.

It is (somewhat) important to how to phrase your inferences from confidence intervals and p-values. For example, the following statement from Gelman and Hill (2007), p. 20 is wrong (The authors say so themsevlves):

> The hypothesis of whether a parameter is positive is directly assessed via its confidence interval. If both ends of the 95% confidence interval exceed zero, then we are at least 95% sure (under the assumptions of the model) that the parameter is positive.

What would be better to say is that, **if all model assumptions are satisfied, then in repeated applications the 95% confidence intervals will include the true value 95% of the time.** This is *not* however the same as a precise statement about the *probability* of the parameter being poistive. You need more and stronger assumptions for that. (You need a prior distribution asumtption, which we will not discuss)

Two important things to note:

1. All inferences are conditional on your model being true!
2. With classical methods, we are making statements about the probability of a true parameter being in a range *given infinite repititions of our experiment*, not about the probability of a parameter having a certain value. These are two very different concepts (At least to Statisticians)

Many more insights can be gleamed for the American Statistics Association's statment on p-values

Asymptotics

For a lot of more complicated situations, where we need different assumptions and get different $se(\hat{\beta})$ estimators for example, it quickly becomes impossible to figure out the shape of the distribution of $t_{\hat{\beta}} = \frac{\hat{\beta}-\beta_0}{se(\hat{\beta})} \sim t_{n-k}$ is. In such cases one can often still make an argument what that distribution looks like if the sample becomes

big. Because very often much of the randomness becomes closer and closer to a normally distributed random variable as the sample grows. In such cases, we can develop asymptotic tests. For example, see Wooldridge (2010) p.53 for asymptotic normality. The central assumptions here are:

1. Random sampling
2. Population orthogonality assumption $E[x'u] = 0$
3. rank E(x'x) = K (which means no perfect co-linearity)
4. $E[uu'|X] = \sigma^2 I$ (homoscedasticity)

Most of the problems here are finding asymptotically converging distributions for the cases where we don't want to assume homoscedasticity. Which is basically consistent with assuming that there are unmeasured influences left, but we have some idea about how the affect the behavior of the error term. We will talk more about that in a later session.

References

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevelhierarchical Models.* Cambridge University Press.

Goldberger, Arthur Stanley. 1991. *A Course in Econometrics.* Harvard University Press.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.