STATISTICAL INFERENCE AND HYPOTHESIS TESTING

1. Estimators

- With the concept of a random variable we can now define mathematically what we mean by statistical inference. Let Y be some random variable, e.g.

$$Y \sim \text{Normal}(\mu, \sigma).$$

Statistical inference is the process of estimating the parameters (e.g. μ and σ) based on samples of Y AND expressing our uncertainty in these estimates.

- As we have already pointed out many times, in order to estimate parameter θ, we can use the following facts:

  (1) Both μ and σ can be represented as means over the distribution of Y. For example $\theta = E[\theta]$.

  (2) If we have enough samples the sample average should be close to the actual average. That is, $1/N \sum_{i=1}^{N} f(Y_i) \approx E[f(Y)]$. The central limit theorem tells us how accurate this estimate is.

  In general, we will let $\hat{\theta}$ denote an estimator of a parameter θ from a sample if $\hat{\theta}$ is some function of our sample which is meant to approximate θ. For example

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

  is an estimator of μ in the Normal model. It also happens to be the sample mean: $\hat{\mu} = \bar{Y}$.

- $\hat{\theta}$ **vs.** θ **Remember:** the estimator is a function of the data. That is, $\hat{\theta}$ **depends on the specific data we collect** or simulation we run. It is meant to approximation a parameter which μ does not depend on the data and is (in classical statistics) a fixed number. For example, in the instance of a YES/NO survey or election with two candidates, the "true" quantity we are interested in measuring is the fraction of people answering YES to some question. Our estimate, $\hat{q}$, is a variable which depends on the specific subset of the population we sample.

1.1. Sample distribution and standard errors.

- Since $\hat{\theta}$ depends on the data, different replications of our sample will generate different values of $\hat{\theta}$. We can therefore think of $\hat{\theta}$ as a random variable itself. We call the distribution of $\hat{\theta}$ over many replications of our data the sample distribution. I will use replicate to mean different realizations of our data (as opposed to the different samples within the data). The distinction is shown in Figure **??** (left panel). The terminology gets a bit confusing: The sample distribution is the distribution of $\hat{\theta}$ over many replicates, but each replicate involves many samples.

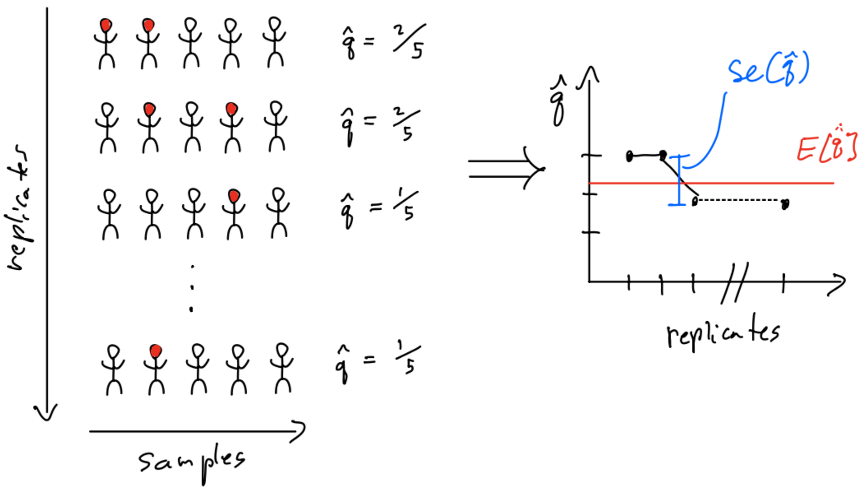

Figure 1. Replicates and samples

Example 1 (sample distribution of normal mean). *Suppose*

$$Y \sim \text{Normal}(\mu, \sigma)$$

Question: What is the sample distribution of $\hat{\mu}$ (our estimate of μ)?

Solution:

$$\hat{\mu} = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

The CLT tell us (informally speaking) that

$$\sum_{i=1}^{N} Y_i \sim \text{Normal}\left(\mu n, n\sigma^2\right)$$

where by $\sim$ we really mean "approximately distributed as". Dividing by N,

$$\hat{\mu} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{N}\right)$$

This assumes σ is known.

- A natural way to quantify the uncertainty in our estimate is the standard deviation of the $\hat{\mu}$ under the sample distribution. We call the resulting quantity the standard error, which is our estimate of the standard deviation of the sample distribution. For the Normal model, if we are estimating the mean and happen to know σ, then

$$\text{se}(\hat{\mu}) = \frac{\sigma}{\sqrt{N}}. \tag{1}$$

  This tells us how much our estimate will vary between different experiments (or surveys/simulations). Importantly, the standard error depends on σ which we may not know!!! Thus, it is common to estimate the standard error using an estimate of σ, $\hat{\sigma}$, leading to an estimator of the standard deviation:

$$\text{se}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{N}}. \tag{2}$$

  It should be clear from the context which one we are talking about: If we are working with data and we don't know what σ is, when we say standard error we mean Equation 2. If we are working with a particular model where we have specified the parameters, we mean Equation 1.

1.2. Bias and consistency.

- There must be some properties we would like the estimator to have. At a minimum, it should be in some way informed by the data, in the sense that having more data should bring our estimate closer to the actual value of the parameter. More precisely, the more data we have (e.g. the larger N) the closer we expect $\hat{\theta}$ to be to the true value θ. Of course, we must define what we mean by "closer" when we are talking about random things. For our purposes we will say $\hat{\theta}$ is consistent if

$$E[\hat{\theta}] \to \theta \quad \text{and } \text{se}(\hat{\theta}) \to 0 \quad \text{as} \quad N \to \infty.$$

  This is saying that as we obtain more and more samples, the sample distribution because more concentrated around θ.

- To see that consistency is not the only property we look for in an estimator, notice that since $\hat{\mu}_1 = \hat{\mu} + 1/N$ is also consistent, yet clearly seems inferior to $\hat{\mu}$. To this end, we say that an estimator $\hat{\theta}$ is unbiased for some N (not just very large N), the average over the sample distribution is equal to the actual value under the model distribution; that is,

$$E[\hat{\theta}] = \theta.$$

  Example 2 (Bias and consistency)**.** *For a normal random variable, define the following estimators of the mean:*

$$\hat{\mu}_2 = \frac{Y_1 + Y_2}{2}$$

  Question: Is $\hat{\mu}_2$ biased and consistent? what is the sample distribution?

  Solution: Note that $\hat{\mu}_2$ has the sample distribution

$$\hat{\mu}_2 \sim \text{Normal}(\mu, \sigma/\sqrt{2})$$

  Example 3 (Normal standard deviation)**.** *Let now consider estimating the standard deviation of a Normal random variable*

$$Y \sim \text{Normal}(\mu, \sigma^2)$$

  Given samples $Y_1, Y_2, \dots, Y_n$, it seems the natural way to estimate σ^2 is using

$$\text{var}(Y) = E[(Y - E[Y])^2] \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

  we will call this estimator $\hat{\sigma}_0^2$. It turns out $\hat{\sigma}_0^2$ is biased and in-fact

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{n}{n-1}\hat{\sigma}_0^2$$

is unbiased. The correction by a factor $n/(n-1)$ is called Bessel correction.

Question: Demonstrate with simulated data that $\hat{\sigma}_0^2$ is biased and $\hat{\sigma}^2$ is not.

1.3. Confidence intervals.

- The idea of the confidence interval is, roughly speaking, to describe the range of values where we think the actual value of θ might reasonable be given some estimate $\hat{\theta}$ and its sample distribution. We will mostly work with the 95% confidence interval, or 95%-CI, which is given by

$$[\hat{\theta} - 1.96\text{se}(\hat{\theta}), \hat{\theta} + 1.96\text{se}(\hat{\theta})] \tag{3}$$

The factors 1.96 in front of the standard errors ensure that 95% of samples from the sample distribution will fall in this range,

Note that these samples from the sample distribution do not have the same distribution as $\hat{\theta}$ over replicates of our data. Said another way, if we draw many samples from our estimate of the sample distribution, their distribution will not be the same as the distribution of $\hat{\theta}$ we would obtain if we ran an experiment many times and estimated $\hat{\theta}$ each time. The correct interpretation of the 95%-CI is as follows: **If we generate many replicates of the data then θ (the true value) will fall in the CI, for 95% of them.**

Technically speaking, is is NOT the case that there is a 95% chance the true value of θ is in the 95%-CI. To understand why, note that the parameter has a 95% chance to be in the interval

$$[\theta - 1.96\text{std}(\theta), \theta + 1.96\text{std}(\theta)] \tag{4}$$

but this is difference from Equation 3, since we have replaced $\hat{\theta}$ with θ. The distinction, which is shown in Figure 2, is important; however, you don't need to get bogged down by the subtle differences in interpretation. For practical purposes, you can pretty much thing of the 95%-CI as the region where the parameter value is likely to be. We provide alternatives ways to think about these intervals when we discuss Bayesian vs. classical statistics.

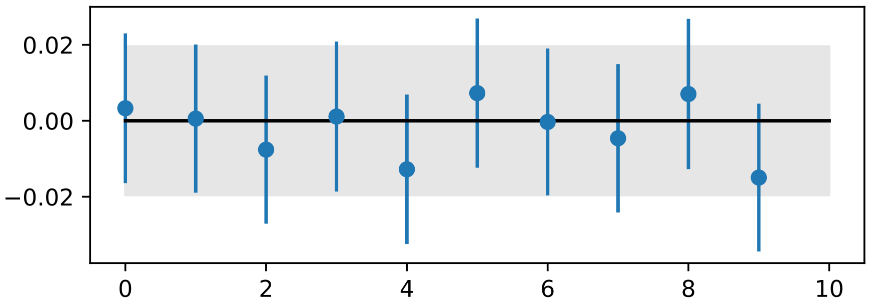

Figure 2. An illustration of the distinction between Equation 4 (gray shaded region) and Equation 3.

Example 4 (Estimating CI)**.** *Imagine we are designing an experiment. Our model is a Normal distribution and from previous experience, we have a ballpark estimate of the standard deviation, which is $\sigma \approx 0.1$.*

Question: Roughly, how many samples do we need to collect to have a 95% chance our estimate is within 0.1 of the actual value of the variable?

Solution: The standard deviation of an estimate based on n samples will have a confidence interval of

$$[\hat{\mu} - 0.196/\sqrt{n}, \hat{\mu} + 0.196/\sqrt{n}]$$

The width of this interval is $2 \times 0.196/\sqrt{n}$. This interval will intersect the true value in 95% of replicates, so we would like it to have a width < 0.2. It follows that we need

$$1.96^2 = 3.8 < n$$

We can test this by running many replicates for each n, as done in the class notebook.

2. Maximum Likelihood

- Sometimes it is quite clear what the estimator for a parameter should be. This is the case for q in the Bernoulli distribution. However, we will find this is not always the case, so it is useful to have a **more systematic way of finding estimators.**

- Recall that the probability distribution for the binomial distribution is

$$p(Y) = \binom{n}{Y} q^Y (1-q)^{n-Y} \tag{5}$$

In statistics, we sometimes call this the likelihood and denoted $P(Y) = L(Y|q)$. The notation here is suggesting that we think of P as a distribution which is conditioned on a particular value of the parameter.

More generally, the likelihood is defined as the probability we say a data set given the parameters. This notation and terminology foreshadows Bayesian thinking, wherein one thinks of the parameter as random variables themselves – more on this later.

- For now, notice that Equation (5) tells us how likely it is to observe k YES among n people surveyed. Then, it seems reasonable that this number should not be very small, since that would mean our survey results are an anomaly. More generally, the larger $L(Y|q)$ is the more likelihood our results are. This suggests one a way to estimate determine q: We can take as our estimate $\hat{q}$ the value which makes $L(Y|q)$ largest. In other words, we are finding the value of q which makes the data the most likely, and we will call this the maximum likelihood estimate.

- You can do this using calculus (if you know how, I suggest you give it a try) to determine that the value of q which makes (5) largest is
$$\hat{q}_{\text{MLE}} = \frac{Y}{n}$$

- For a Normal distribution with mean and variance μ and σ, the MLE estimators are the usual sample mean and standard deviation which we have already been exposed to.

3. Statistical inference for regression model using python

- Having introduced the concepts and terminology of statistical inference, we return to the linear regression model with a single predictor:
$$Y|X \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2).$$
We will sometimes write such a model as
$$Y = \beta_0 + \beta_1 X + \varepsilon$$
where it is assumed that
$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$
Note that this model does not describe the distribution of Y, rather it describes the distribution of Y given X. What we are really saying is that once we are given a value of X, the variation in Y is approximately normal with an X independent variance. The formulas for estimators of β_0 and β_1 have already been given. You can look up formulas for the standard errors in the textbook, we will just have Python give us these. Note that the sample distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are NOT normal, but we will not worry about trying to compute them in this class (at least for now).

- In addition, we want to estimate σ^2. To do so, we can note that
$$Y - (\beta_0 + \beta_1 X) \sim \text{Normal}(0, \sigma^2).$$
Therefore, with our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can compute the residuals of each data point
$$r_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$
An estimate of σ^2 is approximately the sample variance of r_i. We need to account for the fact that we don't know β_0 and β_1 exactly though (just like with Bessels correction), and therefore the unbiased estimator is
$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2.$$
We will refer to the process of computing the coefficient estimates and standard errors as fitting the model.

- In order to perform statistical inference for a linear regression model in Python, we use the `statsmodels` package. `statsmodels` has a function `OLS` which can take two arguments

  - An array `y` – the response variables
  - and a matrix `X` – which contains the predictors.

  Technically, `OLS` will create a regression model of the form
$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + Z$$
where Z is mean zero normal. Notice there is not intercept, instead we have a term $\beta_0 X_0$. We need to trick `OLS` into creating a model with a single predictors AND an intercept by creating a new predictor X_0 which is always one. This is achieved with the line of code

```
> X = sm.add_constant(x)
```

where x is jus the array of our X values (playing the role of X_1 in the equation above). Then

```
> model = sm.OLS(y,X)
```

create a model "object". At this point we haven't actually performed any inference, rather we have created an data structure, called an object, which has all the information about our model and data. The next step is to tell python to fit this model – that is, to infer the parameter values – and store the results. This is done via

```
> results = model.fit().
```

Results contains all the information about the fitted model, including the inferred, or fitted, coefficients, standard errors, as well as many things I will soon discuss. We can print all this out with the line of code

```
> print(results.summary())
```

This will produce something like

```
>                             OLS Regression Results
> ==============================================================================
> Dep. Variable:                      y   R-squared:                       0.612
> Model:                            OLS   Adj. R-squared:                  0.610
> Method:                 Least Squares   F-statistic:                     312.1
> Date:                Thu, 05 Oct 2023   Prob (F-statistic):           1.47e-42
> Time:                        12:43:13   Log-Likelihood:                -519.05
> No. Observations:                 200   AIC:                             1042.
> Df Residuals:                     198   BIC:                             1049.
> Df Model:                           1
> Covariance Type:            nonrobust
> ==============================================================================
>                  coef    std err          t      P>|t|      [0.025      0.975]
> ------------------------------------------------------------------------------
> const          7.0326      0.458     15.360      0.000       6.130       7.935
> x1             0.0475      0.003     17.668      0.000       0.042       0.053
> ==============================================================================
> Omnibus:                        0.531   Durbin-Watson:                   1.935
> Prob(Omnibus):                  0.767   Jarque-Bera (JB):                0.669
> Skew:                          -0.089   Prob(JB):                        0.716
> Kurtosis:                       2.779   Cond. No.                         338.
> ==============================================================================
```

The value of `const` is the intercept β_0 and `x1` is the coefficient of X in our regression model – that is, the slope β_1. You can access the various quantities directly from the results object via the commands in Table 1.

quantity	Command
β_i	`results.params`
$\hat{\sigma}^2$	`results.scale`
se($\hat{\beta}_i$)	`results.bse`

Table 1.

Example 5 (Performing linear regression in statsmodels)**.** *Here we once again consider the some on advertising budgets and sales for a company.*

Question: Fit the data to a linear regression model using `statsmodels` *and compare to the results we got before.*

3.1. Assessing explanatory power with R^2.

- We have so far seen how to estimate and interpret the parameters in the regression model with one predictor). So far, most of the quantities we compute depend heavily on the intrinsic scales of the data. For example, if we are working with height data as our predictor and use units of inches, we will get very different values of $\hat{\beta}_1$ than if we had used feet. The same is true for the other parameters. It is therefore useful to have a summary of the descriptive power of our model in terms of a quantity that does not depend on these scales.

- To find an appropriate metric for accessing the descriptive power of our model, the idea is to compare the variation in Y over all (that is, the marginal variance), to the variation conditioned on X. To this end, we define the correlation coefficient by

$$\rho^2 = 1 - \frac{\mathrm{var}(Y|X)}{\mathrm{var}(Y)} = 1 - \frac{\sigma^2}{\beta_1^2\sigma^2 + \sigma_x^2} \approx R^2 \tag{6}$$

  The quantity called R^2 is simply the estimator of this with all the quantities above replaced by their sample estimates:

$$R^2 = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

- There is a relationship between ρ and the covariance

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_x\sigma_y}.$$

  I'll leave it as an exercise to show this. This provides provides another interpretation of ρ (and hence R^2) which comes from one of the earlier exercises. In particular, if β_1 and β_1' are respectively the regression slopes of Y vs. X and X vs. Y, we can then write ρ as

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_x\sigma_y} = \frac{\text{sign}(\beta_1)\sqrt{\beta_1'\beta_1}\sigma_x\sigma_y}{\sigma_x\sigma_y} = \text{sign}(\beta_1)\sqrt{\beta_1'\beta_1}.$$

Example 6 (Performing linear regression in statsmodels)**.** *Question: Generate simulated data with different values of R^2*

Solution: See python notebook.

4. Hypothesis testing

- In statistics, we might infer parameters not because we are interested in specific values, but rather because we would like to use them to make a decision. For example, in a clinical trial, we might be interested in deciding whether a candidate drug is worth moving forward with. This problem is often framed in terms of hypothesis testing, in which we assign a probability to a particular hypothesis or its converse.

- In rather abstract terms, the basic procedure of hypothesis testing is as follows:

(1) Come up with a null hypothesis. For example, this might be that the mean of some variable is zero. We are interested in determining whether we can rule this hypothesis out.

(2) Compute something called a test statistic, denoted $\hat{T}$, which like any estimator is simply some quantity we compute from our data.

(3) Next, we do a sort of probabilistic thought experiment and ask: What is the chance that we would observe a value of $\hat{T}$ at least as large as the value we measured IF our hypothesis was in-fact true. The result is the p-value.

Example 7 (hypothesis testing for a clinical trial)**.** *Consider the example of a clinical trial. The effect of a drug, denoted Y (e.g. blood pressure is measured in two groups) is measured in two groups. One group is given a placebo, the other (the treatment group) is given a drug. Let $X = 0$ for people in the control group and $X = 2$ for those in the treatment group. For simplicity we will assume that there are $N/2$ people in each group. We can model Y with a regression model*

$$Y|X \sim \text{Normal}(\mu_C(1-X) + \mu_T X, \sigma^2)$$

We will assume σ^2 is know! This greatly simplifies the calculations! *This is just a linear regression model since we could write*

$$\mu_C(1-X) + \mu_T X = \mu_C + (\mu_T - \mu_C)X = \beta_0 + \beta_1 X$$

where

$$\beta_0 = \mu_C$$
$$\beta_1 = \mu_T - \mu_C.$$

We could estimate β_0 and β_1 as we always do in a linear regression model. We could also simply perform inference on the mean and of a Normal distribution within each group to obtain estimators of μ_C and μ_T. For simplicity, let's pretend σ is known for simplicity. This makes things simple, because then the sample distributions are

$$\hat{\mu}_C \sim \text{Normal}\left(\hat{\mu}_C, \frac{\sigma^2}{N/2}\right)$$
$$\hat{\mu}_T \sim \text{Normal}\left(\hat{\mu}_T, \frac{\sigma^2}{N/2}\right).$$

Thus the (estimated) sample distribution of β_1 is

$$\hat{\beta}_1 \sim \text{Normal}\left(\hat{\beta}, \frac{4\sigma^2}{N}\right).$$

In this case, our null hypothesis will be that $\beta_1 = 0$; that is, there is no effect of the drug. As our test statistic, we measure how far β_1 is from zero in standard deviations:

$$\hat{T} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

Remember that since we know σ, $\text{se}(\hat{\beta})$ is known and therefore, from the perspective of the sample distribution, this is just dividing by a constant. Now, let $\hat{\beta}_1^*$ be the random variable representing the measured effect under the null hypothesis. Another way to say this is that $\hat{\beta}_1^*$ represents a measurement of β_1 from a replica generated under the assumption that $\beta_1 = 0$. Therefore, $\hat{\beta}_1^*$ will have a distribution centered at zero and with a standard deviation $\text{se}(\hat{\beta}_1)$. This means the distribution of $\hat{\beta}_1^*$ is nothing but the sample distribution shifted to zero, or

$$\hat{\beta}_1^* \sim \text{Normal}\left(0, \frac{4\sigma^2}{N}\right)$$

At this point we can answer the question posed in step 3: **If the null hypothesis was true, how likely would we be to observe a value of $\hat{T}$ larger than the one we did?** *This is determined by the p-value:*

$$p_v = P(|\hat{T}^*| > |\hat{T}||\hat{T}) \tag{7}$$

where $\hat{T}^*$ is the test statistic computed from $\hat{\beta}_1^*$ and the probability is taken over all the distribution of $\hat{T}^*$, while $\hat{T}$ is given by our data (hence why I use the conditioning notation). p_v, like $\hat{T}$, is a function of the data. See the python notebook were we compute p_v with simulations.

- The above example is very simple because we assume that σ is known and we have only a binary predictor. In reality, the computation of p-values is much more complex, however the principle and interpretation is the same!

- **Interpreting the p-value** If the p-value is very small, then it is highly unlikely we would have observed what we did when the null hypothesis was true. In this case, we can REJECT the null hypothesis as false. Usually some threshold is set for this, and if the p_v is below that threshold we say our result in statistically significant. On the other hand, **if p_v is not small, it does not necessarily mean the null hypothesis is true.** A result is said to be statistically significant if $p_v < 0.05$. Visually, we can see that β_1 is statistically significant exactly if 0 is not contained in the confidence interval!

- **Relationship between p-values and confidence intervals**. The p-value is all about the "tail" of the sample distribution – "tail" usually just means the ends of the distribution. Naturally there is connection between between p-values and confidence intervals, which also measure the width of the sample distribution. To illustrate the connection, we will again assume σ **is known**. Since the the sample distribution can be obtained by shifting the distribution of $\hat{\beta}_1^*$ to $\hat{\beta}_1$, the p-value, p_v, is exactly the chance of being outside the interval $[\hat{\beta}_1 - |\hat{\beta}_1|, \hat{\beta}_1 + |\hat{\beta}_1|]$. Therefore, recalling the interpretation of confidence intervals, $\hat{\beta}_1$ will fall in the confidence interval with probability p_v when the null hypothesis is true. If σ isn't known all this is only approximately true, but intuition is still.

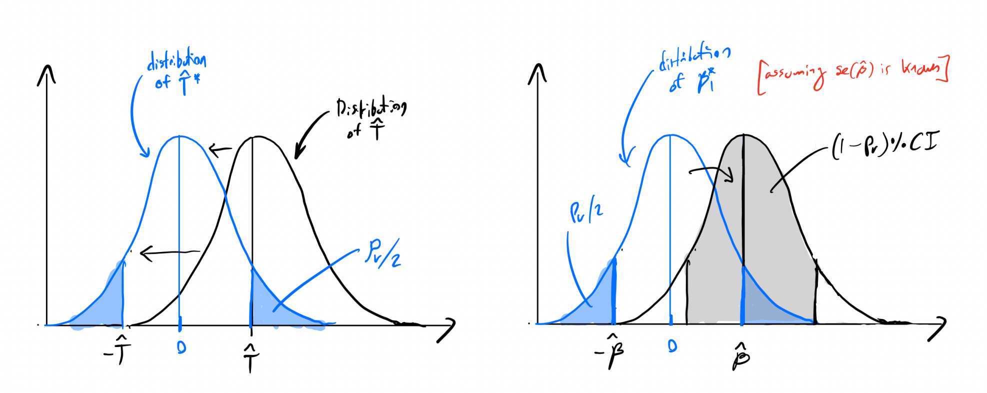
distribution of $\hat{T}^*$
Distribution of $\hat{T}$
$P_v/2$
$-\hat{T}$
0
$\hat{T}$
distribution of β_1^*
[assuming se($\hat{\beta}$) is known]
$(1-P_v)\%$ CI
$P_v/2$
$-\hat{\beta}$
0
$\hat{\beta}$

Figure 3. (left) The (two-sided) p-value and (right) the relationship between p_v and the confidence interval.

References

[1] John Tabak. *Probability and statistics: The science of uncertainty*. Infobase Publishing, 2014.

[2] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning (python version)*, volume 112. Springer, 2013.