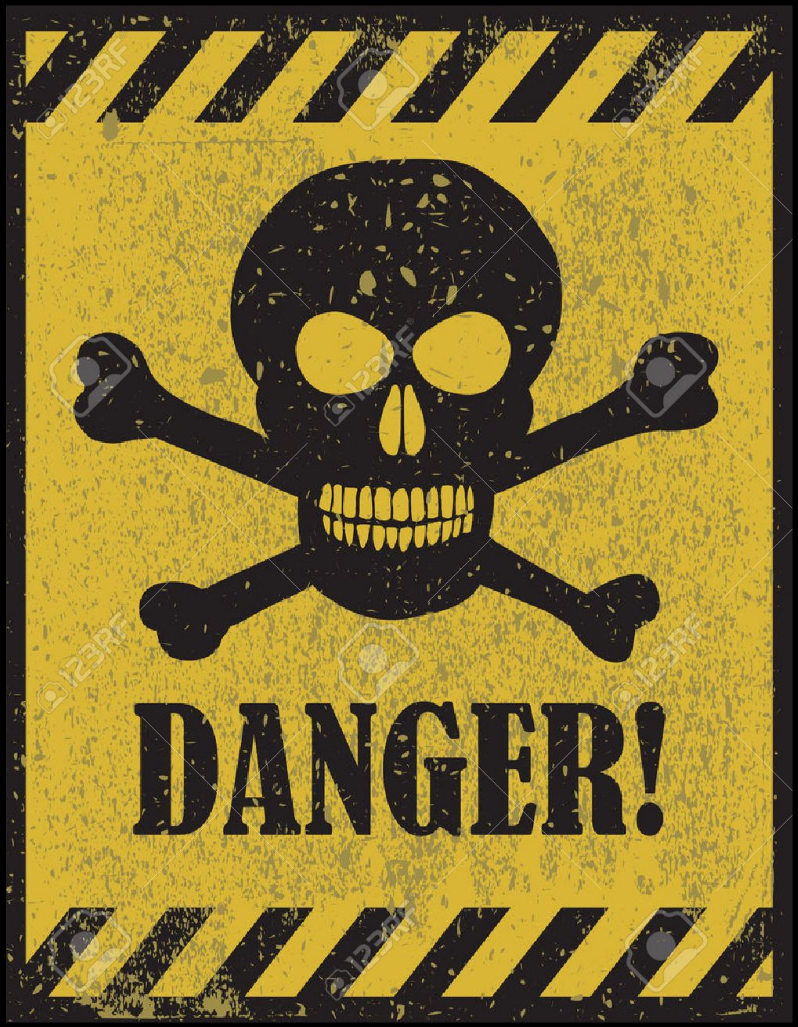
First notions on statistical inference

- So far we have seen how to ESTIMATE parameters from data (e.g. mean)

- We now need ways to assess how precise is the information these estimates are providing

- In other words, we want to quantify UNCERTAINTY around our estimates

- One way we can quantify uncertainty is by looking at the variability of an estimate. We can compute its VARIANCE/SD.



This IS NOT saying that the parameter is a RANDOM VARIABLE, with some values being more likely than others.

In this framework, we assume that there is A VALUE that is the TRUE VALUE OF THE PARAMETER

The reason why we observe VARIABILITY in our ESTIMATE of the parameter is because of SAMPLING.

Example. True model (data generating process):

X: outcome of throwing a “fair” the dice

Possible outcomes (“probability space”): {1,2,3,4,5,6}

Probability distribution: all outcomes have same probability of occurrence: P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6

What is the mean (or expected value) ?

$$\mu = \frac{1}{N} \sum_{\text{All x}} x_i$$

$$\mu = \sum_{\text{All x}} P(x_i) * x_i = (1/6) * 1 + (1/6) * 2 + (1/6) * 3 + (1/6) * 4 + (1/6) * 5 + (1/6) * 6 = 3.5$$

You could also compute the variance and sd: 2.91 ; 1.71

So,

If we RANDOMLY select 500 people to INDEPENDENTLY throw a fair dice, and we record the results, we know we should expect an average of mu = 3.5, with and sigma=1.71

BUT….

- Because we are working with a sample, we might not get these value :
  mu != mu_hat

- mu_hat may change from sample to sample

- The question is: what can we expect from our estimate?:

    - What is the average of mu_hat?
    - What is the sd of mu_hat?

Believe it or not, mu_hat will follow a NORMAL DISTRIBUTION with:

$$mean(\hat{\mu}) = \mu$$

$$sd(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

DEMONSTRATION IN R

Why?

Some properties of mean and variance. **If X1 and X2 are iid with**

$$mean(X_1) = \mu_1$$

$$var(X_1) = \sigma_1^2$$

And

$$mean(X_2) = \mu_2$$

$$var(X_2) = \sigma_2^2$$

$$mean(X_1 + X_2) = mean(X_1) + mean(X_2) = \mu_1 + \mu_2$$

$$mean(a * X_1) = a * mean(X_1) = a * \mu_1$$

$$var(X_1 + X_2) = var(X_1) + var(X_2) = \sigma_1^2 + \sigma_2^2$$

$$var(a * X_1) = a^2 * var(X_1) = a^2 * \sigma_1^2$$

So…

$$mean(\hat{\mu}) = mean(\frac{1}{n}(x_i + \ldots + x_n)) = \frac{1}{n}(mean(x_i) + \ldots + mean(x_n))$$

$$= \frac{1}{n}(\mu + \ldots + \mu) = \frac{n\mu}{n} = \mu$$

And

$$var(\hat{\mu}) = var(\frac{1}{n}(x_i + \ldots + x_n)) = \frac{1}{n^2}(var(x_i) + \ldots + var(x_n))$$

$$= \frac{1}{n^2}(\sigma^2 + \ldots + \sigma^2) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$sd(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$$

The sd of an estimate is ofter referred as the "Standard Error" of the estimator or SE

Implication: The larger the sample, the more precise are our "estimates" but the gains of increasing n are marginally decreasing