Representative samples: samples that faithfully reflect the underlying population.

Samples are slightly dependent because N is not exactly infinity.

$$X_1, X_2, \ldots, X_n, \quad T_n = X_1 + \ldots + X_n$$

T would be the rv of the total number of iPhones (x=1's) in the sample of size n, the distribution of the total is:

$$P(T_n = t) = \frac{\binom{\chi}{t}\binom{N-\chi}{n-t}}{\binom{N}{n}}$$, the "hypergeometric distribution"

Many textbooks have units on "finite population" / "small population" inference where you see a lot of corrections for the population size N. We will ignore it from now on... How? We assume the population is infinite.

Let $N \to \infty$, let $\chi \to \infty$ but $\chi / N = \theta$. This implies

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} Bern(\theta) \quad \text{and} \quad T_n \sim Binomial(n, \theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$$

The infinite population sampling assumption is equivalent to the "process sampling" setting. Imagine a factory producing iPhones and you check a sample of iPhones 1 = defective, 0 = working. You can argue there's no population of size N and no χ, the number of defective iPhones in that population. But θ is well-defined: it's the probability of being defective. And inference for θ is very important. In this case, $X_1, \ldots, X_n \overset{iid}{\sim} Bern(\theta)$.

At this point, we assume iid Bernoullis and only care about the best inference for θ. The iid Bernoullis are called the "data generating process" (DGP) and the DGP produces samples of size n.

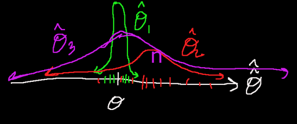
Let's return to point estimation, the first goal of statistical inference, i.e. produce the best single numeric guess of θ, $\hat{\hat{\theta}}$, "statistical estimate"

$$\hat{\hat{\theta}} = \frac{1}{n}(x_1 + \ldots + x_n) \overset{hope}{\approx} \theta$$

$\hat{\hat{\theta}}$ varies dataset-dataset (sample-sample). In one sample of n=5, <1,0,0,1,0> whose average is 0.4; in another x=<1,1,1,0,1> whose average is 0.8. So this point estimation is a realization from the rv called the "statistical estimator":

$$\hat{\theta}_n = \frac{1}{n}(X_1 + \ldots + X_n)$$

The distribution of the estimator is called the "sampling distribution". Its properties are very important.



$$E[\hat{\theta}_n] = E\left[\frac{1}{n}T_n\right] = \frac{1}{n}E[T_n] = \cancel{\frac{1}{n}}(\cancel{n}\theta) = \theta \checkmark$$

The property that the estimator's expectation is the parameter of interest is important. This estimator is said to be "unbiased".

$$Bias[\hat{\theta}] := E[\hat{\theta}] - \theta \overset{if}{=} 0$$

$\neq 0$ if "biased"

What is this expectation "over"? All possible samples of size n.

$$E[\hat{\theta}] = \sum_{\hat{\hat{\theta}} \in \Theta} \hat{\hat{\theta}}\, p_{\hat{\theta}}(\hat{\hat{\theta}}) = \sum_{x_1} \cdots \sum_{x_n} w(x_1, \ldots, x_n)\, p_{\hat{\theta}}(w(x_1, \ldots, x_n))$$

Let's talk about accuracy in general. How "far" is thetahathat from theta? We need to define "far". This is called "loss". We need a "loss" function which takes in thetahathat and θ and provides a non-negative value:

$$\ell : \Theta \times \Theta \longrightarrow [0, \infty) \quad \text{and} \quad \ell(\hat{\hat{\theta}}, \theta) = 0 \text{ iff } \hat{\hat{\theta}} = \theta.$$

Some "popular" loss functions:

$\ell(\hat{\hat{\theta}}, \theta) = |\hat{\hat{\theta}} - \theta|$ absolute loss or L_1 loss

$\ell(\hat{\hat{\theta}}, \theta) = (\hat{\hat{\theta}} - \theta)^2$ squared error loss or L_2 loss (DEFAULT)

$\ell(\hat{\hat{\theta}}, \theta) = (\hat{\hat{\theta}} - \theta)^p$ where p > 0 it's called L_p loss

An important thing you want to know is the mean loss. That's called "risk":

$$R(\hat{\theta}, \theta) = E[\ell(\hat{\theta}, \theta)] \overset{\text{if } \ell = \text{sqd. loss}}{=} MSE[\hat{\theta}, \theta] := E[(\hat{\theta} - \theta)^2]$$

the default risk function is the mean squared error

if $\hat{\theta}$ is unbiased, $E[\hat{\theta}] = \theta$

$$= E[(\hat{\theta} - E[\hat{\theta}])^2] = Var[\hat{\theta}]$$

The MSE of an unbiased estimator is the variance of the estimator.

if $\hat{\theta}$ is biased

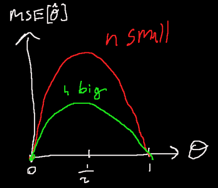
$$= E[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] = \underbrace{E[\hat{\theta}^2]}_{= Var[\hat{\theta}] + E[\hat{\theta}]^2} - 2\theta E[\hat{\theta}] + \theta^2$$

$$= Var[\hat{\theta}] + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 = Var[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2$$

$= Var[\hat{\theta}] + Bias[\hat{\theta}]^2$ bias-variance decomposition of the MSE

Let's look at the MSE of the sample proportion in the iid Bernoulli sampling case.

$$MSE[\hat{\theta}] = Var[\hat{\theta}] = Var\left[\frac{1}{n}T_n\right] = \frac{1}{n^2}Var[T_n] = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}$$



as n increases

You get really good estimation when θ is near 0 or 1. You get worse estimation when θ is close to 0.5.

That's all for point estimation for the time being. Let's move on to goal #3 (theory testing or "hypothesis testing").

Let's say you're trying to convince your friend that aliens exist. You can either

(I) say you're right and leave the burden on your friend to disprove you
(II) you can assume the opposite and prove to your friend that the opposite is not tenable (bogus)

Which way is more intellectually honest? (II)
Which way will win you more believers? (II)
If you have a vested, monetary interest, which way is more honest? (II)
If you are establishing new laws of physics, which way historically has it gone? (II)

An example of the last is a drug company proving their drug works and they can sell it and make a lot of profit.

How do you prove a theory in a world with random data? Note: since the data is random, you can never be absolutely sure your theory is correct. You can only quantify the uncertainty.

In our context, "theories" are called "hypotheses" and they are a mathematical statement about the parameter of interest in the DGP at hand e.g. I think the proportion of defective iPhones is > 1%.