$$P\left(p \in \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]\right) = 1-\alpha$$

$$\Rightarrow CI_{1-\alpha, p} := \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]$$

Def of the "Confidence Interval" (CI).

However there is trouble here.
This quantity has p in it. And p is unknown. It is the whole goal of inference. If you knew p, you would not need to sample! Thus, the CI as written is non-computable! How to fix?

Fixing this is a famous problem, a problem statistician still are working on today. Here's the fix we teach called the 1-prop CI:

$$P\left(p \in \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right]\right) \approx P\left(p \in \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) = 1-\alpha$$

The fix is to replace p with P-hat since P-hat is approx p if n is large (by law of large numbers thm). This usually approximates the true confidence interval well *if* p is not near 0 or 1.

$$\Rightarrow CI_{1-\alpha, p} := \left[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

level of confidence parameter

Example: flip a coin n=100 times and get 61 heads. Make a 95% CI for the true probability of heads (p).

$$CI_{95\%, p} = \left[.61 \pm 2 \sqrt{\frac{.61(1-.61)}{100}}\right] = [.512, .708]$$

What does a 95% CI really mean? What is a valid interpretation?

(1) Before you collect the data, the probability your CI contains p is 95%
(2) If you collect many many data samples and compute a CI for each one, ≈95% of the CI's will contain p

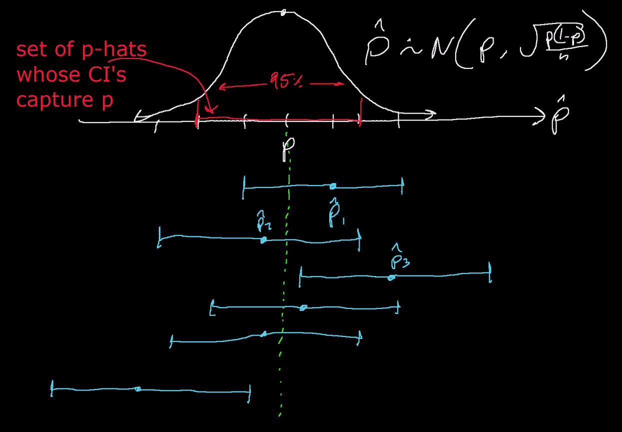
What is an invalid interpretation?

(1) $P(p \in CI_{95\%,p})$ = 95% that is, e.g. the probability p is in [.512, .708] is 95%. In truth, since p is a fixed, constant value, p is in the CI or it's not! Thus the probability is zero or one!

Everyone wants this invalid interpretation to be true! Unfortunately, CI's really mean nothing.

Here's an illustration of the valid interpretations:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$



Statistcal Inference Goal #3: theory testing (hypothesis testing)
For example
* Is the probability of a baby being male 50%? I would like to prove that probability of a baby being male is ≠ 50%.

If you want to prove that p ≠ 50%, we will call this the "alternative hypothesis" (the thing you really do want to prove).

How can we prove the alternative hypothesis?

(1) We can just state by fiat that we're right and demand the world around us to submit proof that we're not right. If they can't then our theory is proven.

This is a weak proof! It's not going to get many followers.

(2) We can play devil's advocate, assume for the moment we're wrong, i.e. assume the opposite of the alternative hypothesis which we call the "null hypothesis" (it is "null" because you seek to "nullify" it). Then, you adduce evidence (data) and hopefully show that this data really is implausible if the null was true. If it is implausible, you "reject the null hypothesis". If it's plausible, you "fail to reject the null" or "retain the null". "Rejecting the null" is kind of like "proof by contradiction".

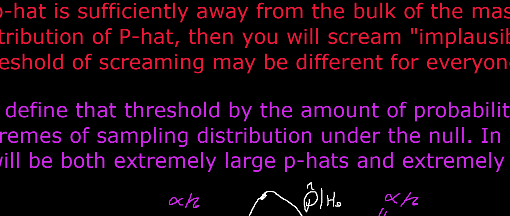
This is a strong proof since everyone sees your intellectual honesty. You're giving the null the best possible chance and rejecting it beyond a doubt. We use this method in statistical inference and call it "hypothesis testing".

We denote the alternative hypothesis by H_a and the null hypothesis by H_0. These two hypotheses will be mathematical statements about the value of the unknown parameter. In our case, the unknown parameter is p. And in our example,

$$H_a: p \neq 0.5, \quad H_0: p = 0.5$$

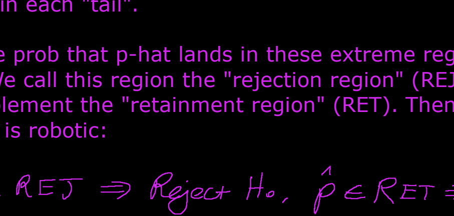
If the null hypothesis is true, then we know what the sampling distribution $\hat{P}$ looks like from the CLT:

$$\hat{p} \mid H_0 \sim N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{n}}^{2}\right)$$



If p-hat is sufficiently away from the bulk of the mass of the distribution of P-hat, then you will scream "implausible". That threshold of screaming may be different for everyone.

We define that threshold by the amount of probability in the extremes of sampling distribution under the null. In this case it will be both extremely large p-hats and extremely small p-hats.



$2.5\% + 2.5\% = 5\% = \alpha$, with $2.5\% = \alpha/2$ and $2.5\% = \alpha/2$

Denoting this probability of extremeness as alpha, then we put alpha/2 in each "tail".

Now, the prob that p-hat lands in these extreme regions is then alpha. We call this region the "rejection region" (REJ) and we call its complement the "retainment region" (RET). Then the decision is robotic:

$$\hat{p} \in REJ \Rightarrow \text{Reject } H_0, \quad \hat{p} \in RET \Rightarrow \text{Retain } H_0$$

A "rejection" is considered a "statistically significant" result. A "retainment" means "there is not sufficient evidence to reject the null".

If the null hypothesis was true, what is the probability you reject it (incorrectly)? Alpha! They thing you control. This is called a Type I error. You control the type I error!

The scientific default for alpha = 5%. This goes back to a paper that RA Fisher wrote in the 1930's. Many journals want 1%.

Let's return to our example. We collect birth data on n = 345 babies and 169 of them were male. Run the hypothesis test at alpha = 5% => z = 2.

$$H_a: p \neq 0.5, \quad H_0: p = 0.5 = p_0$$

$$RET = \left[p_0 \pm z \sqrt{\frac{p_0(1-p_0)}{n}}\right] \quad \text{general formula}$$

$$= \left[0.5 \pm 2 \sqrt{\frac{0.5(1-0.5)}{345}}\right]$$

$$= [.446, .554]$$

$$\hat{p} = \frac{169}{345} = 0.48 \in RET \Rightarrow \text{Retain } H_0$$

There is not sufficient evidence to suggest the prob. of male births is different from 50-50. The sample proportion of 169/345 = 0.48 is plausible given the 50-50 null assumption.