**Variance** (Sample variance: s2 or Population variance: σ2):

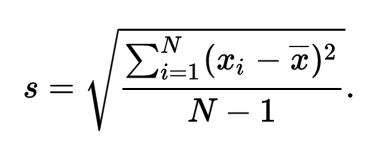
The variance of a random variable X is the expected value of the squared deviation from the mean of X, μ = E[X]

Var(X) = E[(X-μ)2]

= E[X2] – E[X]2

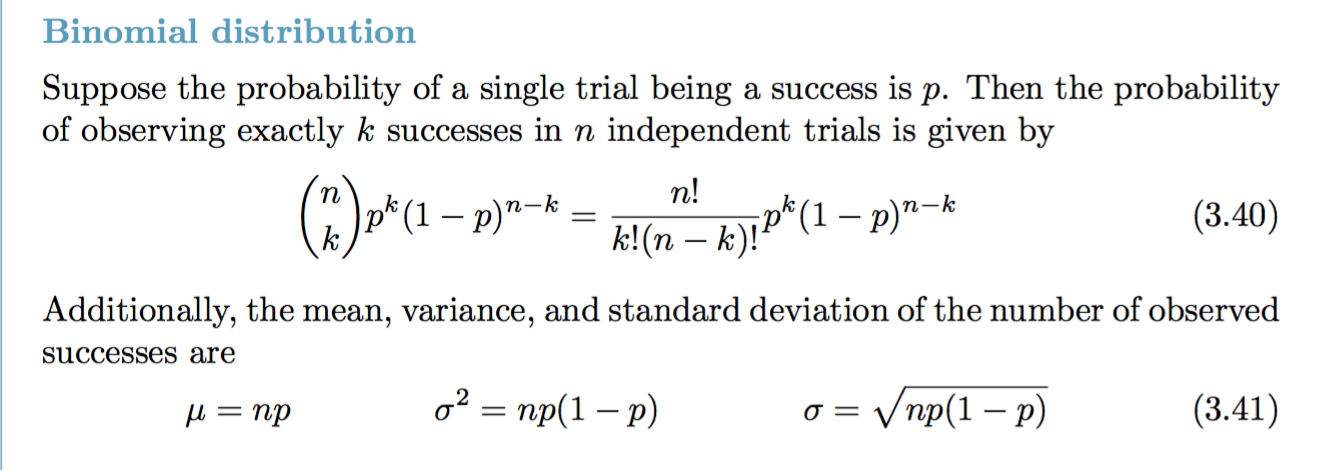
Var(aX) = a2 Var(X)

**Standard Deviation** (s or σ): how variable the population is



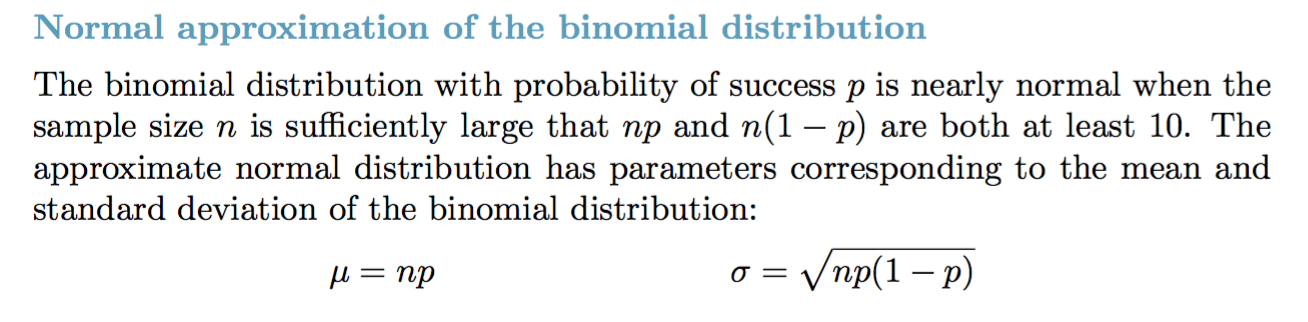
**Standard error** (s/√n): how variable averages of random samples of size n from the population are.

**Binominal mass function:**

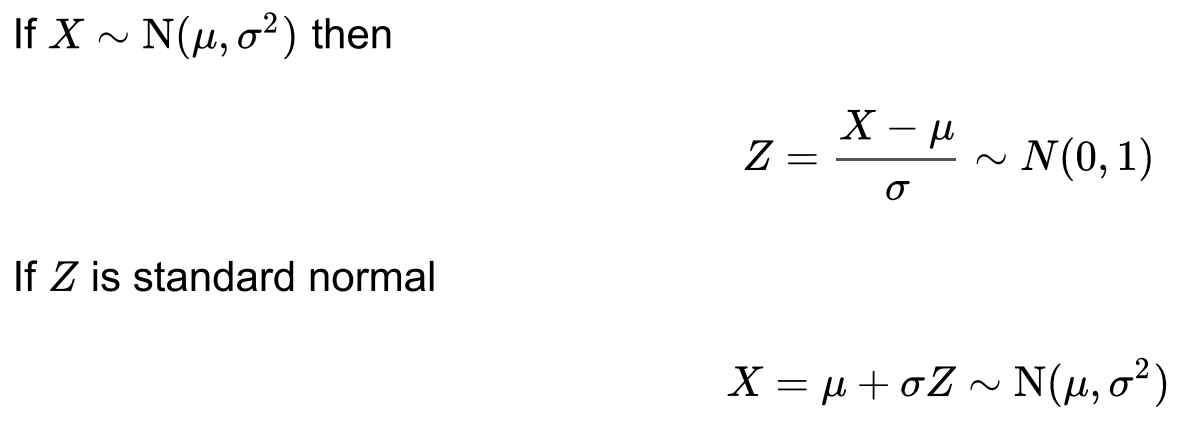
****

binomial conditions:

1. the trials are independent
2. the number of trials, n is fixed
3. each trial outcome can be classified as a success or failure
4. the probability of a success, p, is the same for each trial

****

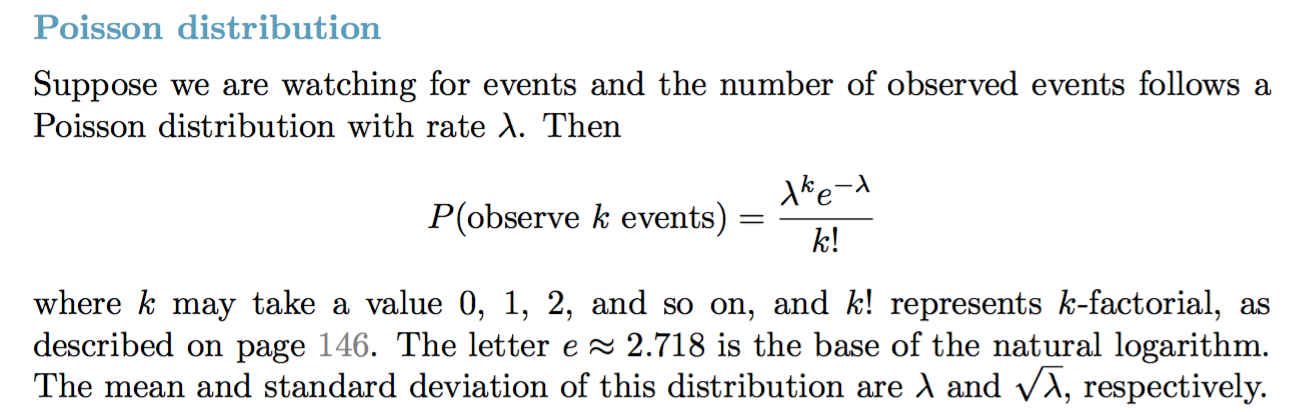
**Standard Normal distribution:**

****

* 68%, 95%, 99% with respect to 1σ, 2σ, 3σ.
* Z score of 1.28, 1.645, 1.96, 2.33 with respect to 90th, 95th, 97.5th, and 99th percentile of the standard normal distribution.

**Poisson distribution:**

Useful for estimating the number of events in a large population over a unit of time.



**T tests**

Hypothesis tests in statistics when you want to compare means.

Compare a sample mean to a hypothesized or target value using a one-sample t-test.

Compare the means of two groups with a two-sample t-test.

The calculations behind t-values compare your sample means to the null hypothesis and incorporates both the sample size and the variability in the data.

A t-value of 0 indicates that the sample results exactly equal to the null hypothesis. As the difference between the sample data and the null hypothesis increase, the absolute value of the t-value increases.

by itself, t-value doesn’t really tell anything, therefore, we need t-distribution to explain.

**t-distributions**

different t-distributions exist for every sample size. A specific t-distribution is defined by its degrees of freedom(DOF).

t-distributions assume that you draw repeated random samples from a population where the null hypothesis is true. You place the t-value from your study in the t-distribution to determine how consistent your results are with the null hypothesis.

eg.: a t-distribution that has a 20 DOF, which corresponds to a sample size of 21 in a one-sample t-test.

Goal: decide whether the t-value is unusual enough to warrant rejection of the null hypothesis. To do that, we need to calculate the probability (**p values**). (use t value and t distribution to calculate probabilities)

* high P values: your data are likely with a true null
* low P values: your data are unlikely with a true null.

If the probability is low enough, we can conclude that the effect observed in our sample is inconsistent with the null hypothesis.

Type I error: Rejecting a null hypothesis when it’s true. (α) -------*false positive*

Type II error: Accepting a null hypothesis when it’s false. (β) -----*false negative*

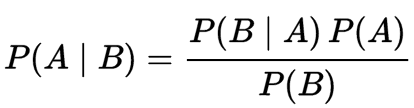
Power: 1-β

**What is the Significance Level (Alpha)?**

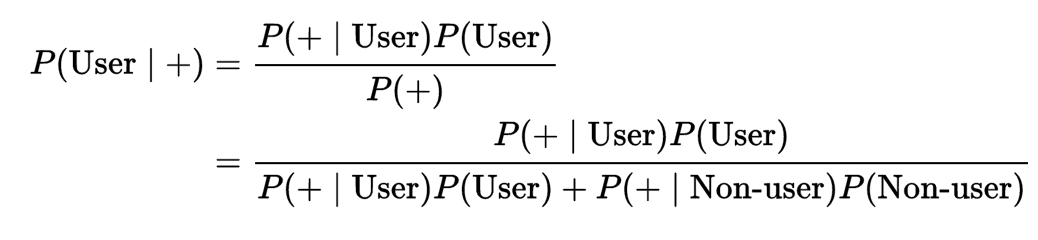
alpha is the probability of rejecting the null hypothesis when it is true.

eg. a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

**Bayes’ Rule:**



Drug test example:



Sensitivity = P (+|D) #you want this to be high

Specificity = P (-|Dc) #you want this to be high

Prevalence of disease = P(D)

Negative predictive value = P (Dc|-)

Positive predictive value = P(D|+)

IID random variable: Independent and identical distributed random variable.

* Observational Study vs. Experiment

*Observational*: based on existing phenomenon, categorize and observe for result (cannot exclude other elements that may affect the result, no definite causal relationship)

*Experiment*: Impose designed action to one group and another control group, compare results. (random assignment)

*CORRELATION DOES NOT IMPLY CAUSATION*

* Mean vs. Median and distribution shapes: (median marks the 50%)

**Left skewed distribution: mean < median (more NO. samples on right)**

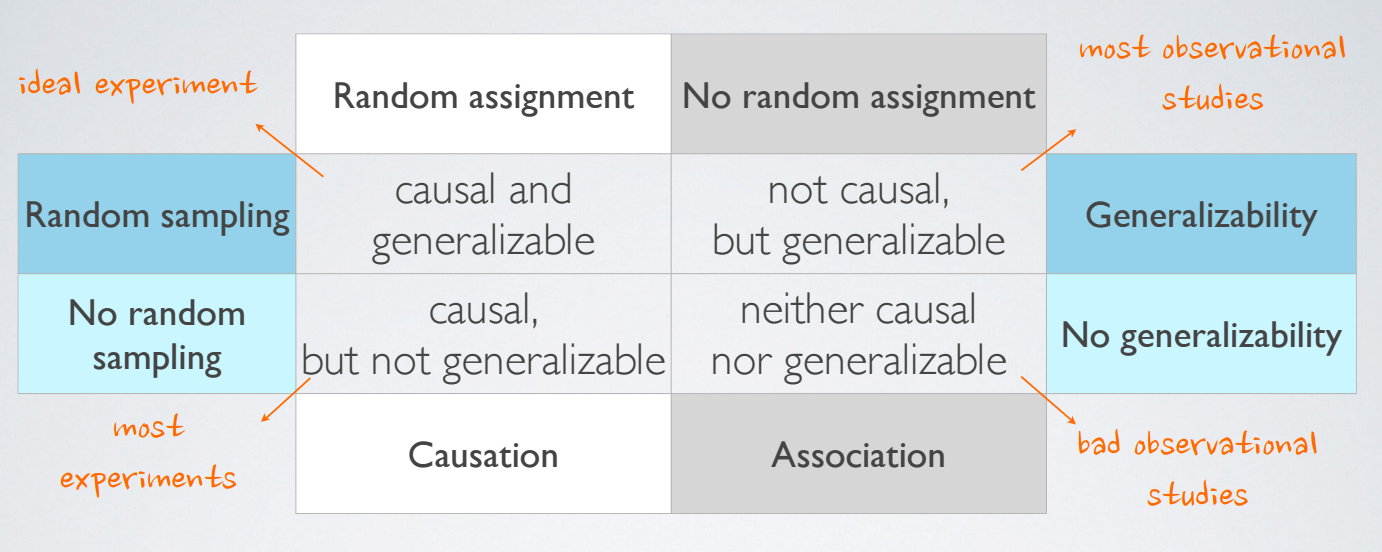
**Symmetric distribution: mean = median**

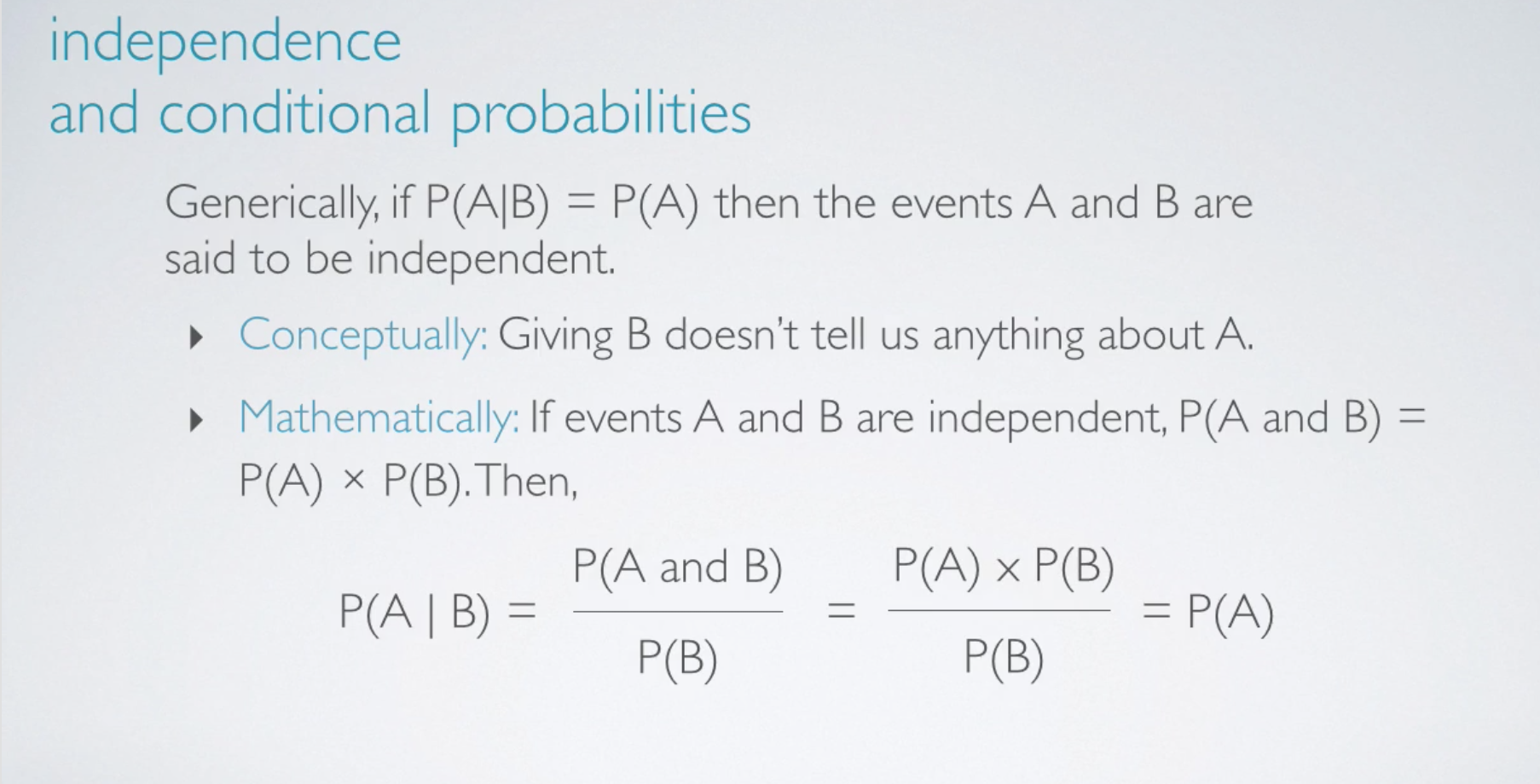
**Right skewed distribution: mean > median (more NO. samples on left)**

* Principles of experimental design
  + control
  + randomize
  + replicate
  + block

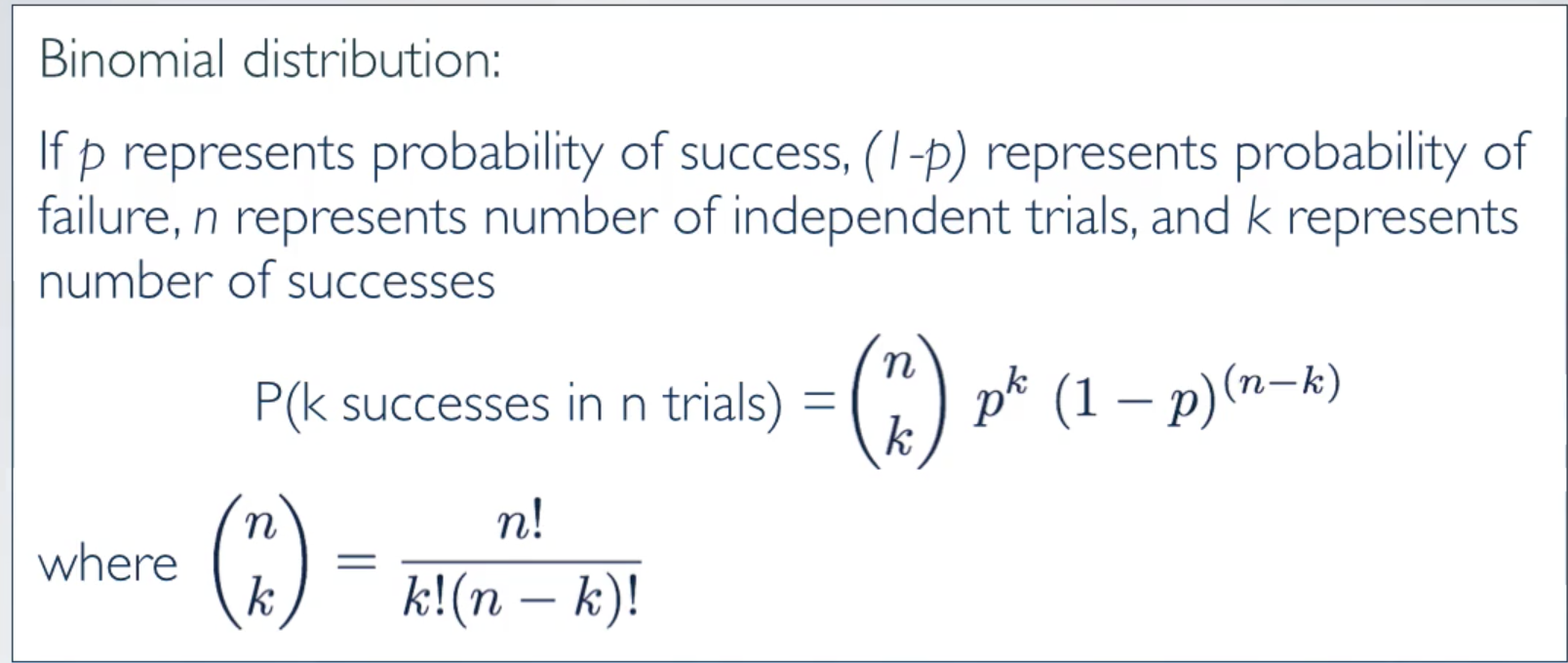
Blocking variables: characters that come with the experimental units that we want to control.

Explanatory variables: also called factors, are conditions we can impose on our experimental units.





* Posterior Probability: P(hypothesis|data), it tells us the probability of a hypothesis, given the data we observed. #a good prior help, a bad prior hurt;
* P(data|hypothesis): P value.



When number of test cases (n) increase, the shape of binomial distribution will look closer to normal distribution. (Rule: success-failure rule: A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution. np >=10; n(1-p)>=10)

If success-failure rule doesn’t apply, adjust the mean value -0.5 when calculating Z score.

Notes on R:

1. read in RData: load(‘selected\_nzes2011.Rdata’)
2. check all variable name: names(selected\_nzes2011)
3. search for a fragment of the name:

grep(‘singlefav’, names(selected\_nzes2011), value = TRUE)

1. tabulating data to see how many times each level:

selected\_nzes2011 %>%

group\_by(jpartyvote) %>%

summarise(count=n())

1. filter function:

selected\_nzes2011 %>%

filter(jpartyvote != ‘Don’t know’) %>%

group\_by(jpartyvote) %>%

summarise(count=n())

1. filter out is.na:

selected\_nzes2011 %>%

filter(!is.na(X\_singlefav)) %>%

group\_by(X\_singlefav) %>%

summarise(count=n())

1. creating a new variable:

selected\_nzes2011 <- selected\_nzes2011 %>%

mutate(sameparty = ifelse(jpartyvote == X\_singlefav, ‘same’, ‘different’))

1. Query data basics:

str(selected\_nzes2011$jnzflik)

str(selected\_nzes2011$jage)

LESSON 2

* **Central Limit Theorem(CLT)**:

The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size

Conditions:

1. independence (n<10% population, random sampling, no replacement)
2. sample size (n>30 for skewed distribution)

* Z score (Z\*) = (mean – sample mean)/S.E.
* **Confidence Interval** (68%, ±1SE); (95%, ±2SE); (99.7%, ±3SE);

The sample mean ± a margin of error:

* Accuracy: CI whether contains the true population parameters
* Precision: width of an interval (CI)

*Under the same sample size, accuracy goes up leads to lower precision*

* Increase sample size, accuracy and precision both goes up.
* ME: Margin of error