A REVIEW OF BASIC STATISTICAL CONCEPTS AND NOTATION

A note on prerequisites

In this set of notes, I will walk through some basic statistical concepts as a quick review.

In this class, I will assume you have learned these concepts in your introductory statistics course(s). In particular, I will assume you know normal distribution. If you have forgotten about these basics, now is the time to review them.

SAMPLE, POPULATION, AND RANDOM VARIABLES

Sample, population, and random variables

A **population** is all individuals/subjects we are interested in.

A **sample** is a selected subset of the population.

There are two general aspects of **statistical inference**—reflecting the interplay between a sample and a population:

- Inductive: Using values observed on individuals in the sample to estimate/infer relevant parameters of the population.
- Deductive: Making probabilistic statements about the sample using information we know or have learned about the population.

In statistical inference, we often focus on **random variables**: a **random variable** is simply some quantity that we measure on individuals in the sample or population.

Example. A clinical trail.

Suppose we want to design a clinical trial to test the effectiveness a blood pressure drug:

- In this case, the **population** will be all potential users of the drug.
- A **sample** will be the people we recruit for the trial.
- One **random variable** can be the blood-pressure measurement before the trial starts.
- We may have a few more random variables that record the blood-pressure measurements at multiple time-points: after a month, a year, etc.
- In a clinical trial, we also often collect information on additional variables (called **covariates**), such as body weight, BMI, etc. These quantities are random variables too.

Factor/categorical variables

Technically, factor/categorical variables are not random variables since they do not take numerical values, but usually we can code the outcome of a factor variable using one or more 0/1 valued indicator variables. The indicator variables are random variables.

For example, in the blood pressure example, we may want to record whether someone is taking any hypertension drugs. We can create an indicator variable which takes a value of 1 if a person is taking some hypertension drug, and 0 otherwise.

When we fit regression models, indicator variables are often used for coding factor variables. In general, if a factor variable has f possible values, we need $f-1$ indicator variables to code its outcomes.

Distribution of a random variable

When we refer to a variable/measurement as a random variable, it usually indicates that we are interested in its **distribution** in the population. In other words, we want to ask 1) **What are all the possible values the random variable can take?** and 2) **How frequent does it take each value?**

A discrete random variable can take finite or countably infinite many values. We can use a table to summarize its distribution: we list all possible values and probabilities (relative frequencies) corresponding to each value.

A continuous random variable can take any value on an interval (the endpoints can be infinity). We can use a density curve to summarize its distribution.

For many commonly-used random variables, their distributions are summarized using a probably mass function (for discrete random variables) or probability density functions (for continuous random variables).

Notation and convention

We often use a capital letter (e.g., Y) to denote a random variable and a small letter (y), to denote the particular value of the variable observed in a sample.

For example, in the blood-pressure drug example, we can use Y to denote blood pressure, y to denote a particular value. Then when we write $Y = y$, it means “the blood pressure is y”.

To list measurements on all individuals in a sample, we can use a subscript to index the individuals: $y_1, \dots, y_n$.

$Y_i = y_i$ means “the blood-pressure measurement of individual i is y_i”.

Example. Binomial random variable and binomial distribution

Suppose we toss a fair coin 3 times and the outcome of each toss is independent of (not affected by) all other tosses. Let Y record the total number of heads we see out of the three tosses, then Y is a binomial random variable. Y can take four possible values, 0, 1, 2, or 3, with probabilities $1/8, 3/8, 3/8, 1/8$ (why?).

We can also summarize the distribution of Y using a table:

Binomial distribution

The coin-tossing example is a special case of the binomial random variable and binomial distribution.

The general binomial distribution with parameters (n, p) describes the distribution of the total number of successes out of n independent trials with probability of success for each trial being p $(0 < p < 1)$.

Each trial can be a coin toss, a success/failure of a test, whether a voter voted for a certain candidate, and so on.

The following probably mass function (p.m.f.) summarizes the distribution of a binomial random variable with parameter (n, p):

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1, \dots, n$$

The probability mass function summarizes all possible values of the random variable and the corresponding probabilities of each value.

The probability mass function of a discrete random variable

The p.m.f. is effectively a compact way to summarize a probability table: all possible values and corresponding probabilities.

Furthermore, the p.m.f. summarizes the probability tables for an entire family of distributions: for example, with each different set of parameters (n, p), the p.m.f.

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,\ldots,n$$

corresponds to a different binomial distribution.

Probability functions in R

In R, we can use

```
dbniom(y, n, p)
```

to compute the probability that a binomial random variable with parameters (n, p) takes a value of y. (You need to specify the values of n and p first.)

```
dbinom(0:n, n, p)
```

will give the entire probability table for a binomial distribution with parameters (n, p).

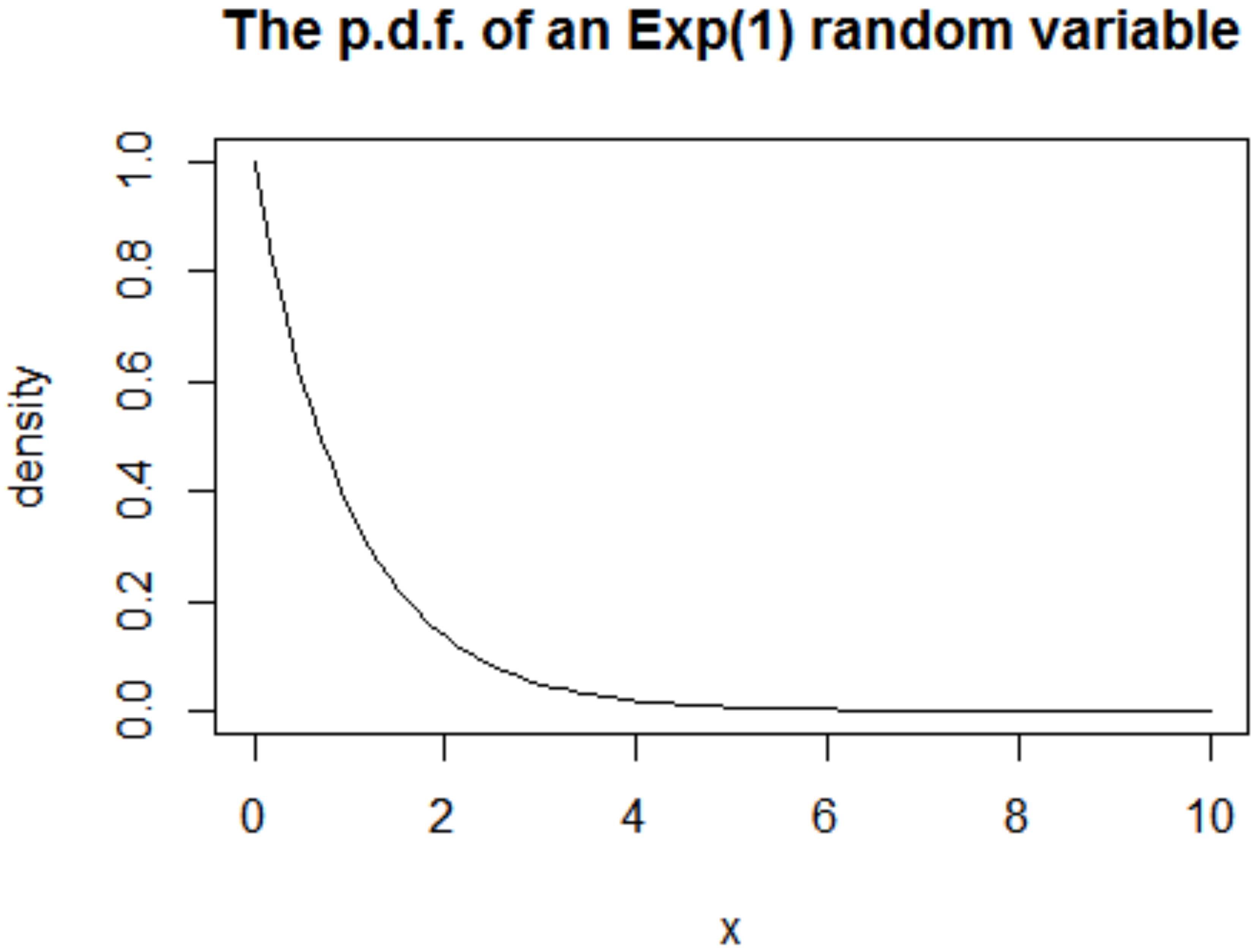
R also provide functions for computing probabilities other commonly used discrete distributions such as: `dpois`, `dgeom`, and so on.

Continuous random variable and probability density

The distribution of a continuous random variable can be summarized by a density curve.

For example, the density curve (probability density function, p.d.f.) of an exponential random variable (with mean 1) is shown on next page.

To compute the probability for a set/region (called an event in statistics) defined by the continuous random random variable, we simply find out the **area under the density curve** for the region. Mathematically, that means we need to integrate the density function over the region.


The p.d.f. of an Exp(1) random variable
density
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
x

Example. Probability calculation for a continuous random variable

Suppose the lifetime of a light bulb can be modeled by an exponential random variable with mean 500 days. What is the probability that the light bulb will last longer than a year (365 days)?

If $f(x)$ is the p.d.f. an exponential random variable with mean 500, then the answer is $\int_{365}^{\infty} f(x)dx$.

In R, we can use the function pexp to compute this probability:

```
pexp(365, rate=1/500, lower.tail=FALSE)
## [1] 0.481909
```

Note:

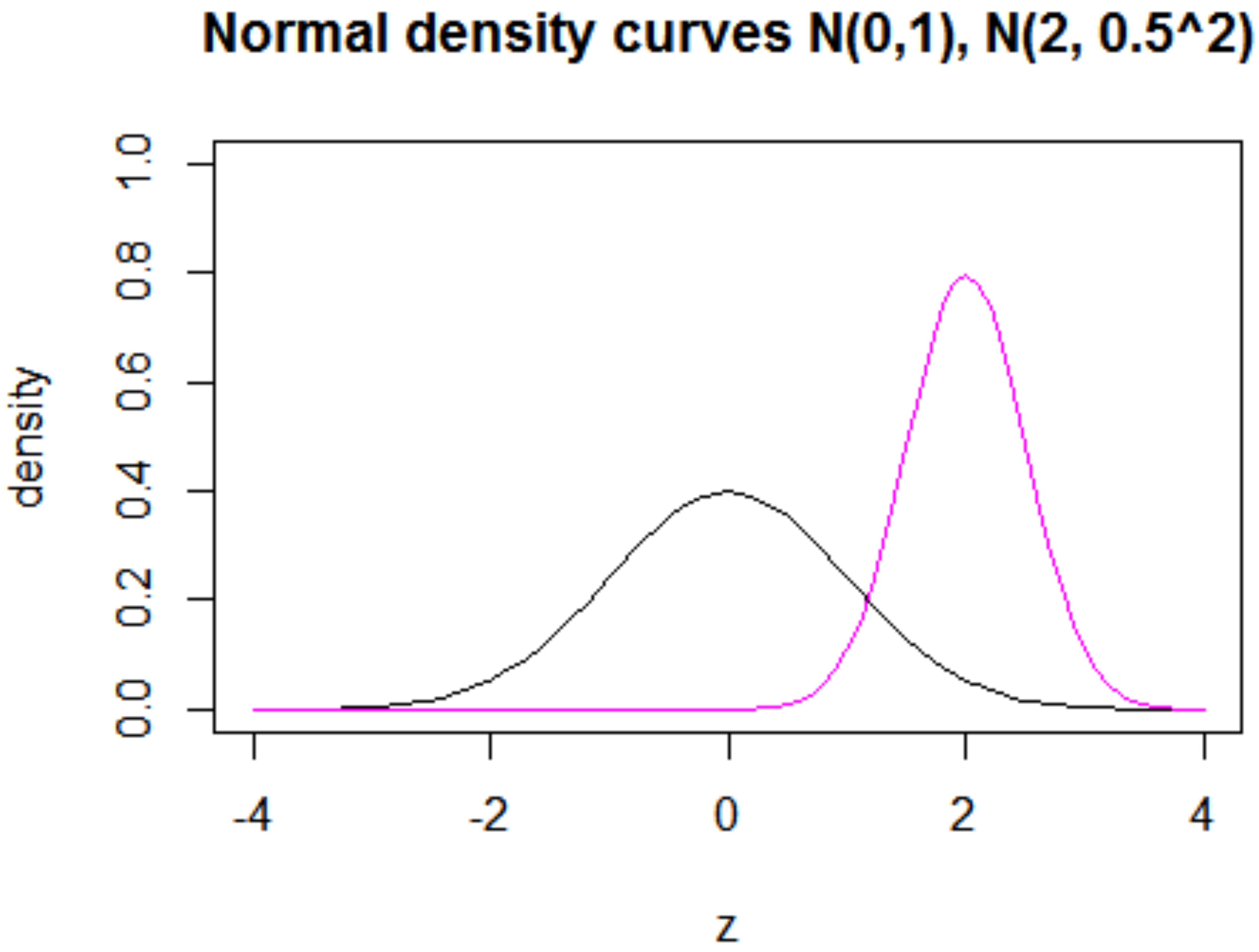
- In pexp, the distribution is specified by the rate parameter, which is $1/\mathrm{mean}$.
- The option $\mathrm{lower.tail} = \mathrm{FALSE}$ instructs R to compute the area under the density curve for the region to the right of 365 (by default, pexp computes the area under the density curve to the left of the value).

Example. Normal distribution.

Normal distribution is the most commonly used model for continuous distributions. It is often used to model the distribution of a sample mean.

The density curve of a normal distribution is the famous “bell-shaped” curve. On next page, we show the density curves (p.d.f.) for $N(0,1)$ and $N(2, 0.5^2)$. $N(2, 0.5^2)$ refers to a normal distribution with mean 2 and standard deviation 0.5. The R code for generating this plot is given below:

```
x = seq(-4, 4, 0.1);
plot(x, dnorm(x), main="Normal density curves N(0,1), N(2, 0.5^2)", type="l",
col="black",
     xlab ="z", ylab="density",
     ylim =c(0,1));
lines(x, dnorm(x, 2, sd=0.5), col="magenta")
```



Normal density curves N(0,1), N(2, 0.5^2)

Probability function for a normal random variable in R

We can use `pnorm` to compute probabilities for a normal random variable.

Suppose $Z \sim N(0,1)$ is a normal variable with mean 0 and variance 1:

Then $\Pr(a < Z < b)$ can be computed as

```
pnorm(b) - pnorm(a)
```

If $X \sim N(\mu, \sigma^2)$, $\Pr(a < X < b)$ can be computed as

```
pnorm(b, mean=mu, sd=sigma)- pnorm(b, mean=mu, sd=sigma);
```

(You have to specify the values of a, b, mu and $sigma$.)

Note that when calling `pnorm`, the `sd` argument is the standard deviation—the square root of the variance—of the normal distribution in question.

Exercise

Suppose $X \sim N(\mu = 2, \sigma^2 = 5)$, use R to compute

- $Pr(X > 5)$
- $Pr(X < 0)$
- $Pr(0 < X \leq 2)$

R functions for p.d.f. and c.d.f.

R provides probability density functions for many commonly continuous distributions: `dnorm`, `dunif`, `dexp`, `dgamma`, …

R also provides probability functions (a.k.a., cumulative distribution functions, c.d.f.) for these distributions: `pnorm`, `punif`, `pexp`, `pgamma`, …

The density functions can be used for plotting the density curves.

The probability functions can be used for computing probabilities. Conceptually, for continuous random variables/continuous distributions, the probability functions are the integrals of the corresponding density functions.

SAMPLE STATISTICS

Sample statistics

Recall that a sample is a subset drawn from a population.

If we measure some quantity on each individual in a sample of size n, that will give rise to a collection of n random variables: $Y_1, \dots, Y_n$, one for each individual in the sample.

Any mathematical function of $(Y_1, \dots, Y_n)$ is called a ***sample statistic***. For example, the sample mean

$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n) = \frac{1}{n}\sum_{1}^{n} Y_i \,,$$

and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

are sample statistics.

Example. An election poll.

Suppose we want to conduct an election poll for Oregon's governor race:

- The population will the all eligible voters.
- The sample will be the subset of individuals we randomly select from the population (i.e., all eligible voters).
- The factor variable that we record will be which candidate an individual intend to vote for.
- Our interest in the poll is often the proportion of voters in the population that will vote for a candidate. That proportion is a **population parameter**.
- We can estimate that parameter using the sample mean (of the indicator variable for voting that candidate). The sample mean is a **sample statistic**.

Random sample and the i.i.d. assumption

When the sample is **randomly** drawn from the population, we often assume that for any quantity we measure, the resulting random variables $Y_1, \dots, Y_n$ are i.i.d.—independent and identically distributed.

The i.i.d. assumption greatly simplifies probabilistic calculations for sample statistics.

Without the i.i.d. assumption, we may have to find out the joint distribution of $(Y_1, \dots, Y_n)$ to do any probability calculations for the sample statistics.

Sample mean and population mean

Suppose we measured some quantity (e.g., blood pressure) in a sample and $y_1, y_2, \ldots, y_n$ are the resulting measurements (random variables). We often estimate the population mean (denoted by μ) of the quantity, by the sample mean,

$$\hat{\mu} = \overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

The sample mean $\overline{y}$ is a statistic: it is a random variable, and each time we draw a different sample, its value will change.

The population mean μ is a parameter: its value is fixed, though often unknown in practice.

When we use the sample mean to estimate the population mean, we say the sample mean is an **estimator** of the population mean.

Sample variance and population variance

The population variance σ^2 is a population parameter, so it is fixed (not random), but its value is usually unknown to us.

In practice, we can estimate the population variance by the sample variance. Suppose $Y_1, \dots, Y_n$ are an i.i.d. with variance σ^2. We can estimate σ^2 by the sample variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

The sample variance, S^2, is a statistic: it is a random and its value varies from sample to sample. In other words, S^2 has a distribution over all random samples (of size n) drawn from the population.

For normal i.i.d. data $Y_1, \dots Y_n \sim N(\mu, \sigma^2)$,

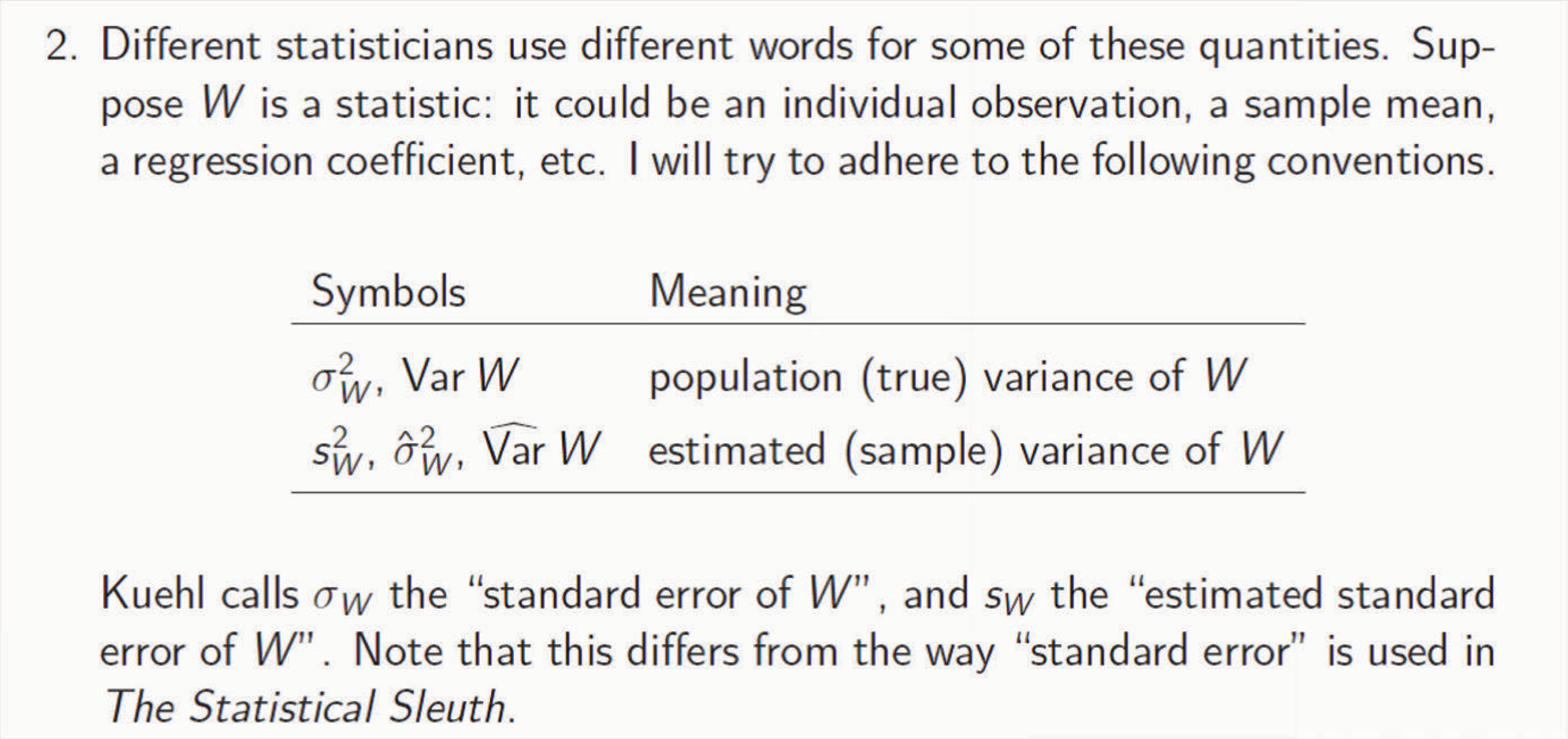
$$(n-1)S^2/\sigma^2 \sim \chi^2(n-1).$$

Conventions

1. Population parameters versus statistics:

- μ and σ^2 are parameters (unknown constants associated with a probability distribution),
- $\overline{y}$ and s^2 are statistics (quantities computed from observed data; random variables).

Conventions

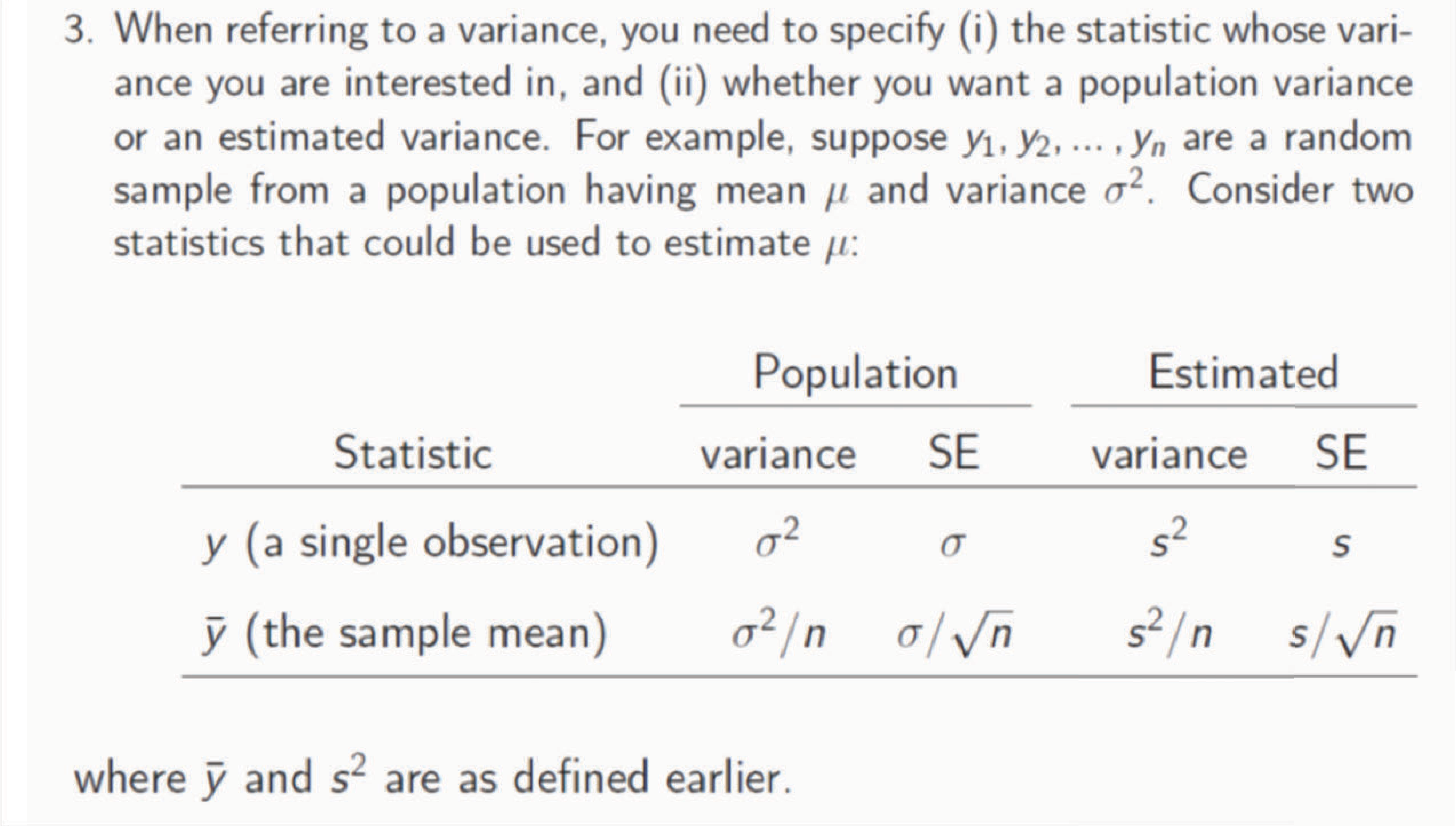


2. Different statisticians use different words for some of these quantities. Suppose W is a statistic: it could be an individual observation, a sample mean, a regression coefficient, etc. I will try to adhere to the following conventions.

Symbols	Meaning
σ^2_W, $\operatorname{Var} W$	population (true) variance of W
s^2_W, $\hat\sigma^2_W$, $\widehat{\operatorname{Var}}\, W$	estimated (sample) variance of W

Kuehl calls σ_W the "standard error of W", and s_W the "estimated standard error of W". Note that this differs from the way "standard error" is used in *The Statistical Sleuth*.

Conventions



3. When referring to a variance, you need to specify (i) the statistic whose variance you are interested in, and (ii) whether you want a population variance or an estimated variance. For example, suppose $y_1, y_2, \dots, y_n$ are a random sample from a population having mean μ and variance σ^2. Consider two statistics that could be used to estimate μ:

	Population		Estimated	
Statistic	variance	SE	variance	SE
y (a single observation)	σ^2	σ	s^2	s
$\bar{y}$ (the sample mean)	σ^2/n	$\sigma/\sqrt{n}$	s^2/n	$s/\sqrt{n}$

where $\bar{y}$ and s^2 are as defined earlier.

Mean and variance for linear combination of random variables

The mean of expected value is a linear operation.

In other words, for linear combination of random variables, e.g.: $Y_1 - 2Y_2 + 4Y_3^2$ the expected value can be taken each summand at a time:

$$E(Y_1 - 2Y_2 + 4Y_3^2) = E(Y_1) - 2E(Y_2) + 4E(Y_3^2)$$

Note that $E(Y_3^2) \neq (E(Y_3))^2$!

For variance, we can take variance term by term **if all terms are independent**, but we have to square the coefficients.

Suppose Y_1, Y_2, Y_3 are independent (which implies Y_1, Y_2 and Y_3^2 are also independent), then:

$$Var(Y_1 - 2Y_2 + 4Y_3^2) = Var(Y_1) + (-2)^2 Var(Y_2) + 4^2 Var(Y_3^2)$$

Note that $Var(Y_3^2) \neq (Var(Y_3))^2$!

Exercise

In this class, what we use the most are linear combinations of independent random variables.

Assume $Y_1, \dots, Y_n$ are independent. Compute the mean and the variance of the following quantities:

- $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $(Y_1 + Y_2)/2 - Y_3$

THE LAW OF LARGE NUMBERS AND THE CENTRAL LIMIT THEOREM

The distribution mean/population mean/expected value

The distribution mean/population mean/expected value (these three terms are interchangeable) of a discrete random variable, Y,

$$E(Y) = \sum_{y} y \cdot p(y),$$

is the weighted average of all possible values it can take, with the weights being the probabilities of each value.

For a continuous random variable,

$$E(Y) = \int y \cdot p(y) dy.$$

We often use the letter μ to denote the population mean to emphasize that it is a parameter (fixed, not random, but maybe unknown to us).

The strong law of large numbers (SLLN)

Let $X_1, \dots, X_n$ be i.i.d. and all have the same distribution as X. Suppose that the expected value of X is finite. Then $\overline{X}_n$ converges almost surely to $\mu = E(X)$.

The SLLN gives us some idea about what the mean μ means in practice: it means if we draw a large i.i.d. sample from the distribution, **the sample mean (statistic) should approach the population mean (parameter)**.

In a coin-tossing example, if we know the probability of head is 0.5, then the expected value (population mean) is $\mu = 0.5$ for the indicator variable for getting head.

This 0.5 means if we toss a coin repeatedly a large number of times (a long-run interpretation) or toss a large number of identical coins (a parallel universe interpretation), then the average number of head (the sample mean) we see should approach 0.5 as the sample size tends to infinity.

The central limit theorem (CLT)

Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean μ and finite variance σ^2. Then as n tends to infinity, the random variable

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

converges in distribution to a standard normal distribution:

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \to Z \sim N(0,1)$$

Sometimes people say that in a large i.i.d. sample, the sample mean is approximately normally distributed.

The CLT says if our interest is mainly in sample means, as is often the case in experimental design setting, we can focus on normal distributions. The CLT is the reason why the normal distributions are the most important distributions in the world.

Distribution of a single random variable versus distribution of a sample statistic

One thing to note is that the distribution of a single random variable Y_i can be very different from the distribution of a sample statistic, such as $\overline{Y}$.

The CLT says that in an i.i.d. sample $Y_1, \dots, Y_n$, it dose not matter what the distribution of each individual Y_i is, the distribution of $\overline{Y}$ is always approximately normal. That's why the CLT is such an important and profound theorem.

The individual Y_i can be discrete (such as binomial, Poisson, …) or continuous (such exponential, uniform, …). It can have any distribution (as long as it has finite mean and variance. The sample mean will always be approximately normal in a large sample.

A sample size of 30 is usually enough for the CLT to apply. If the individual distribution is pretty symmetric, the normal approximation will work well for even smaller sample size.