Parameters and statistics

- Parameters characterize a probability distribution of a population. Ex. μ and σ^2 for the Gaussian distribution; p and N for the Binomial distribution.
- A sample is a subset drawn from the population. A sample is **random** when every element of the population has the same probability of being part of the sample. Henceforth, all samples are assumed to be random. All the results and techniques depend on the sample being random.
- Any quantity or value (like mean, median, variance, proportion, quantile etc.) computed from the elements of the sample is called a **sample statistic** or just **statistic**.
- The main goal of statistical inference is to infer population parameters from sample statistics.
- In practice samples are rarely, if ever, random. Also, we do not have a sample distribution. Most often we have just one sample.
- To that extent the estimates of population parameters from sample statistics are not accurate or true.

Sampling distribution

- The sampling distribution of a statistic, say mean (μ), means the probability distribution of the mean of random samples of size n drawn from the population typically with replacement.
- Since it is a probability distribution we are looking at infinitely many such samples when we are referring to the parameters of the sampling distribution.
- So the sampling distribution is a theoretical construct.
- In simulations, N, the number of samples, only approximates the true sampling distribution when $N \to \infty$ or when it is large enough.
- Typically, large enough can vary depending on the population distribution. But very often $N \geq 30$ for most statistics that we are interested in.

Notation

- SDX will represent the sampling distribution of the statistic X.
- Ex. SDM - sampling distribution of mean; generally SDσ - sampling distribution of std-dev.; SDr - sampling distribution of the Pearson correlation coefficient.
- Population values: greek letters. Ex. μ (mean), σ (std-dev), σ^2 (variance), ρ (Pearson correlation coefficient).
- Estimates for for a sample: greek or latin letter with a cap. Ex. $\hat{\mu}$, $\hat{\sigma}$, $\hat{p}$.
- Values for a sample: RV name with bar. Ex. $\bar{X}$ etc.
- Sampling distribution values: greek with a subscript sd. Ex. μ_{sd}, σ_{sd}.
- N (capital N) - no. of samples.
- n (small n) - size of a single sample.

Example 1

- Three tiles in a bag with values 1, 2, 3. Randomly draw 2 with replacement.
- Outcomes:

No.	1	2	3	4	5	6	7	8	9
T1	1	1	1	2	2	2	3	3	3
T2	1	2	3	1	2	3	1	2	3
Mean	1	1.5	2	1.5	2	2.5	2	2.5	3

Example 1

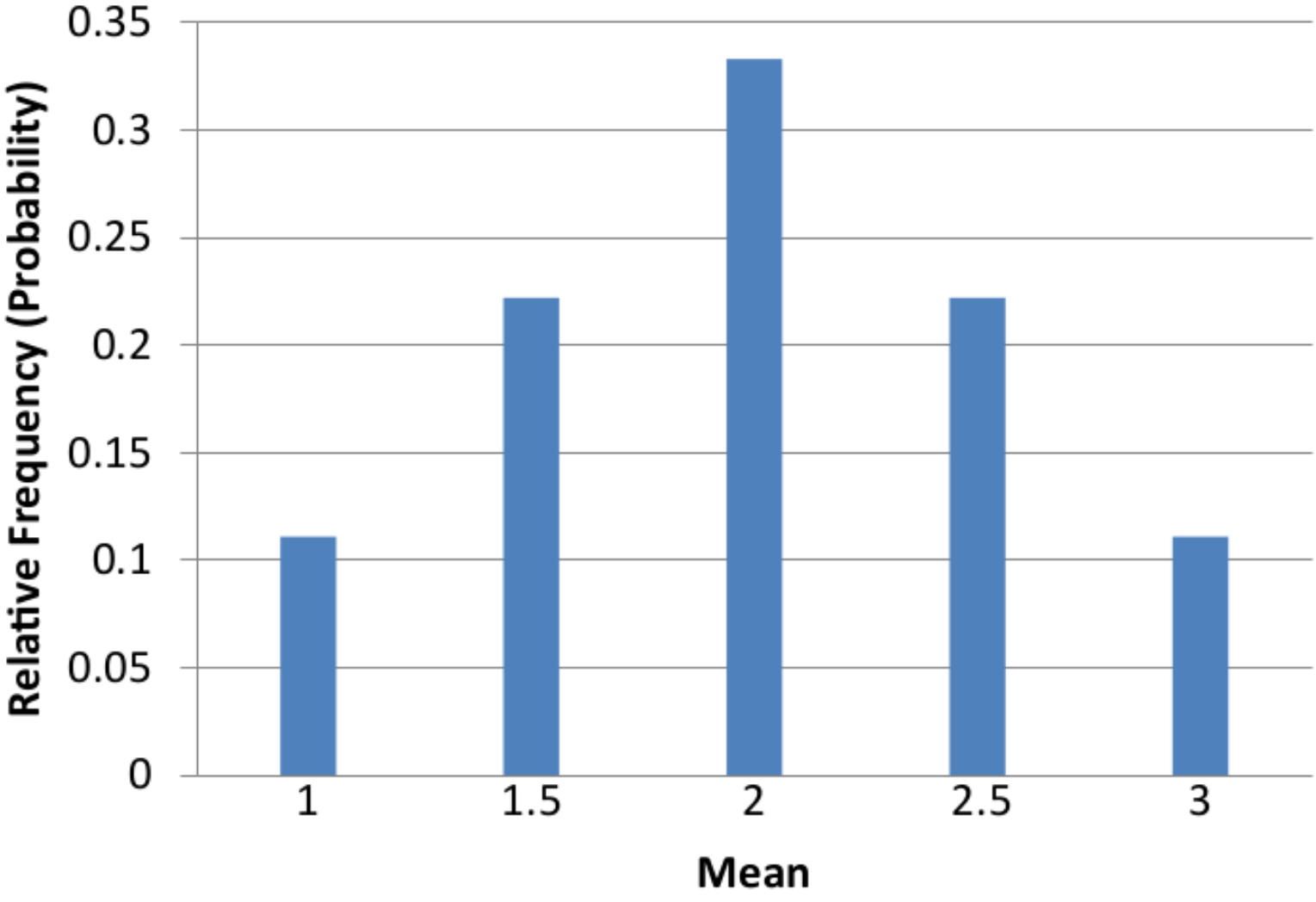
- Three tiles in a bag with values 1, 2, 3. Randomly draw 2 with replacement.
- Outcomes:

No.	1	2	3	4	5	6	7	8	9
T1	1	1	1	2	2	2	3	3	3
T2	1	2	3	1	2	3	1	2	3
Mean	1	1.5	2	1.5	2	2.5	2	2.5	3

- Frequencies:

Mean	1	1.5	2	2.5	3
Freq	1	2	3	2	1

Example 1: histogram



Example 1: alternative

- Keep drawing 2 tiles with replacement N times for large N.
- Draw a relative frequency distribution for the means.
- As $N \to \infty$ we should get the earlier histogram.

Central Limit Theorem (CLT)

First an enabling definition.

Definition 1 (Convergence in distribution or weak convergence)

A sequence $X_1, X_2, \ldots$ of real valued RVs converges in distribtion (written $\xrightarrow{d}$) to RV X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ at which F is continuous. F_n and F are cumulative distribution functions of X_n and X respectively. $\triangleleft$

Definition 2 (Central Limit Theorem (CLT))

Let $\{X_1, X_2 \ldots\}$ be a sequence of iid RVs with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then as $n \to \infty$ the RVs $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0, \sigma)$ distribution. That is:

$$\sqrt{n}(S_n - \mu) \xrightarrow{d} N(0, \sigma)$$

where $S_n = \frac{X_1 + \ldots + X_n}{n}$ is the sample mean. $\triangleleft$

CLT for us - 1

Let $X_1, \ldots, X_n$ be a random sample from any distribution with mean μ and variance $\sigma^2 < \infty$. If n is *sufficiently large* then:

- the sample mean $\hat{\mu}$ follows an approximate normal distribution
- with $E[\bar{X}] = \hat{\mu} = \mu$
- an variance $\sigma^2(\bar{X}) = \hat{\sigma^2} = \frac{\sigma^2}{n}$

Equivalently, $\bar{X} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$ or $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d}$ as $n \to \infty$

CLT for us - 2

What is sufficiently large?

- If the population distribution is symmetric, unimodal or continuous then a sample size as small as $n = 4$ or $n = 5$ is enough.
- If the distribution is skewed then $n \geq 30$ can be adequate.
- If the distribution is extremely skewed then $n \geq 50$ or even $n \geq 100$ may be needed.

Nature of sampling distributions

- The SDX distribution depends on what X is. For sum, mean, proportion the sampling distribution is normal for sufficiently large n.
- For variance, it is a χ^2 distribution.
- More generally, CLT says that a sum of many independent and identically distributed (iid) random variables (or alternatively, random variables with specific types of dependence) will tend to be distributed according to one of a small set of attractor distributions (normal is the most common one).

Application of sampling distributions

- Behaviour of sample mean - for example probability for bounds.
- Test claims. Our major use case.

Example 1: CLT appln

- Let $p(x) = \frac{3}{2}x^2$, sample size $n = 15$. For $-1 < x < 1$ what is probability sample mean falls between $-\frac{2}{5}$ and $\frac{1}{5}$?
- Using definition (needs integration) $\mu = 0$, $\sigma^2 = \frac{3}{5}$.
- From CLT $\hat{\mu} = \mu = 0$ $\hat{\sigma^2} = \frac{\frac{3}{5}}{15} = \frac{1}{25}$.
- We need $P(-\frac{2}{5} < \bar{X} < \frac{1}{5}) = P(-2 < Z < 1)$ that is $0.8413 - 0.0228 = 0.8186$

Example 2: CLT appln

- Waiting time for ith person in a food queue, X_i. Assume claim is average waiting time is 2mins.
- Distribution is a negative exponential distribution (point Poisson process).
- You take a sample of 36 persons and find average waiting time is 3.2mins.
- Is the earlier estimate reasonable? What is the probability that you will obtain a sample mean ≥ 3.2.
- Using CLT $\hat{\mu} = 2$ and $\hat{\sigma}^2 = \frac{4}{36} = \frac{1}{9}$.
- $P(\hat{X} > 3.2) = P(Z > 3.6) = 0.0002$. So, highly unlikely.