LECTURE 8

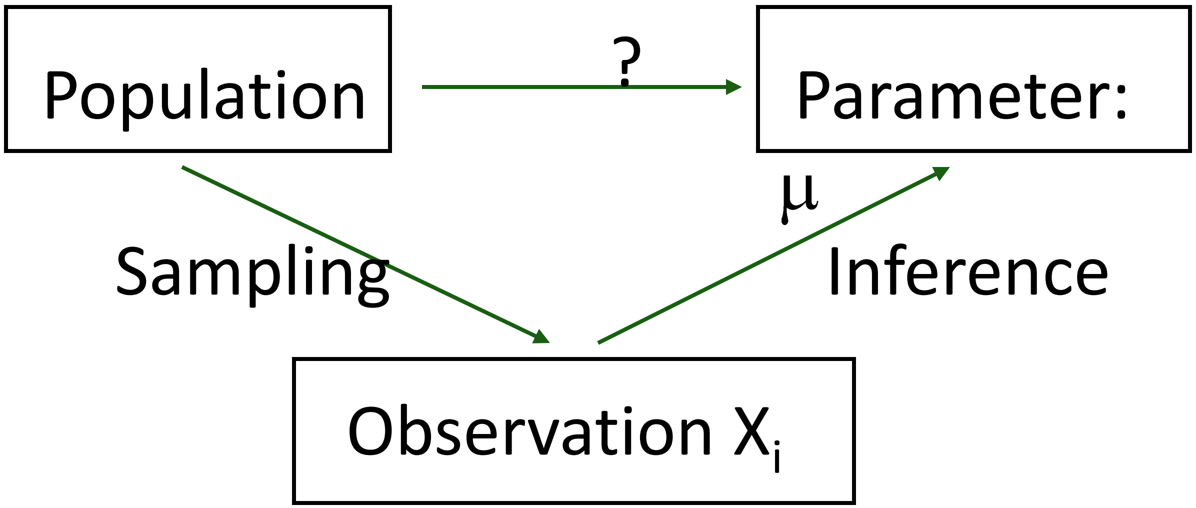
SAMPLING DISTRIBUTION

INFERENCE

- In real life calculating parameters of populations is usually impossible because populations are very large. Rather than investigating the whole population, we take a sample, calculate a **statistic** related to the **parameter** of interest, and make an inference.

- Inferential statistics allow the researcher to come to conclusions about a population on the basis of descriptive statistics about a sample.

INFERENCE WITH A SINGLE OBSERVATION



- Each observation X_i in a random sample is a representative of unobserved variables in population
- How different would this observation be if we took a different random sample?

STATISTIC

- Let $X_1, X_2,\ldots,X_n$ be a r.s. of size n from a population and let $T(x_1,x_2,\ldots,x_n)$ be a function which does not depend on any unknown parameters. Then, the r.v. or a random vector $Y=T(X_1, X_2,\ldots,X_n)$ is called a ***statistic***.

- The ***sample mean*** is the arithmetic average of the values in a r.s.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The ***sample variance*** is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

- The ***sample standard deviation*** is the statistic defined by S.

SAMPLING DISTRIBUTION

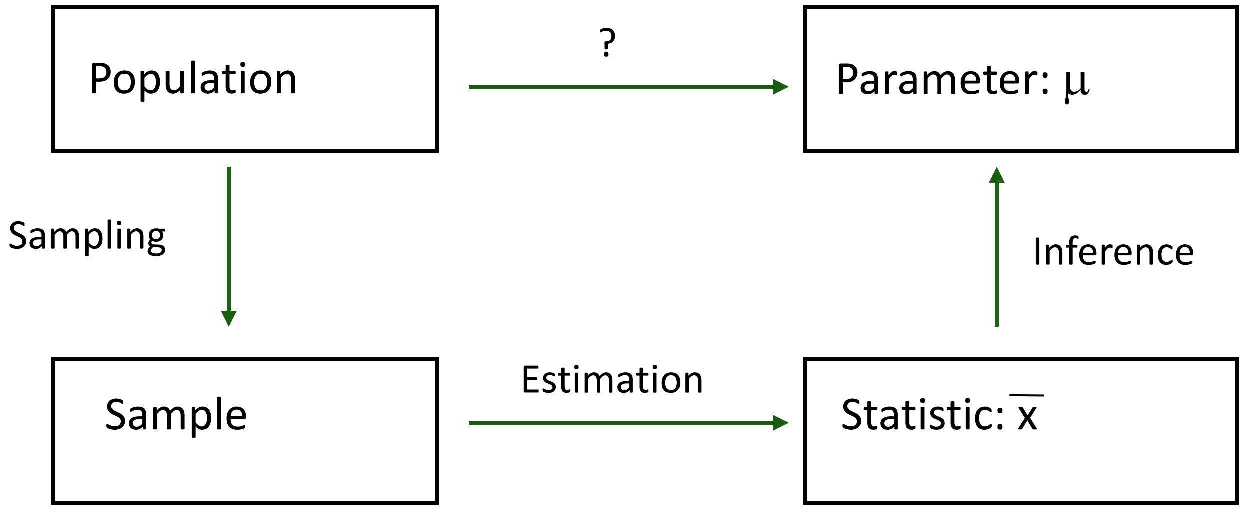
- A statistic is also a random variable. Its distribution depends on the distribution of the random sample and the form of the function $Y=T(X_1, X_2,\ldots,X_n)$.

- The probability distribution of a statistic *Y* is called the ***sampling distribution*** of *Y*.

- A sampling distribution is a distribution of a statistic over all possible samples.
- To get a sampling distribution,
  1. Take a sample of size N (a given number like 5, 10, or 1000) from a population
  2. Compute the statistic (e.g., the mean) and record it.
  3. Repeat 1 and 2 a lot (infinitely for large pops).
  4. Plot the resulting ***sampling distribution,*** a distribution of a statistic over repeated samples.

The method we will employ on the ***rules of probability*** and the ***laws of expected value and variance*** to derive the sampling distribution.

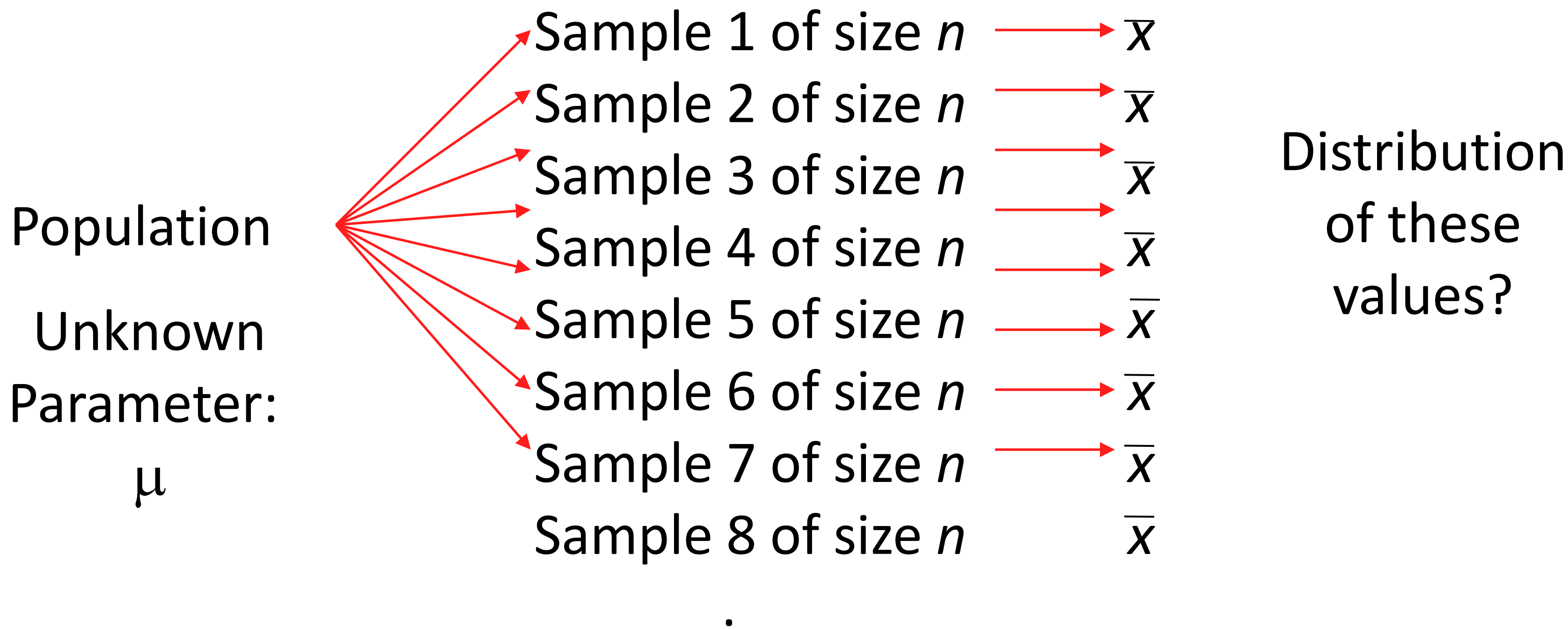
Example: Inference with Sample Mean



- Sample mean is our estimate of population mean
- How much would the sample mean change if we took a different sample?
- Key to this question: **Sampling Distribution** of $\overline{x}$

SAMPLING DISTRIBUTION OF SAMPLE MEAN

- Model assumption: our observations x_i are sampled from a population with mean μ and variance σ^2



Example

- A fair **die** is thrown infinitely many times,with the random variable X = # of spots on any throw.

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

- The probability distribution of X is:

$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

…and the mean and variance are calculated as well:

$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

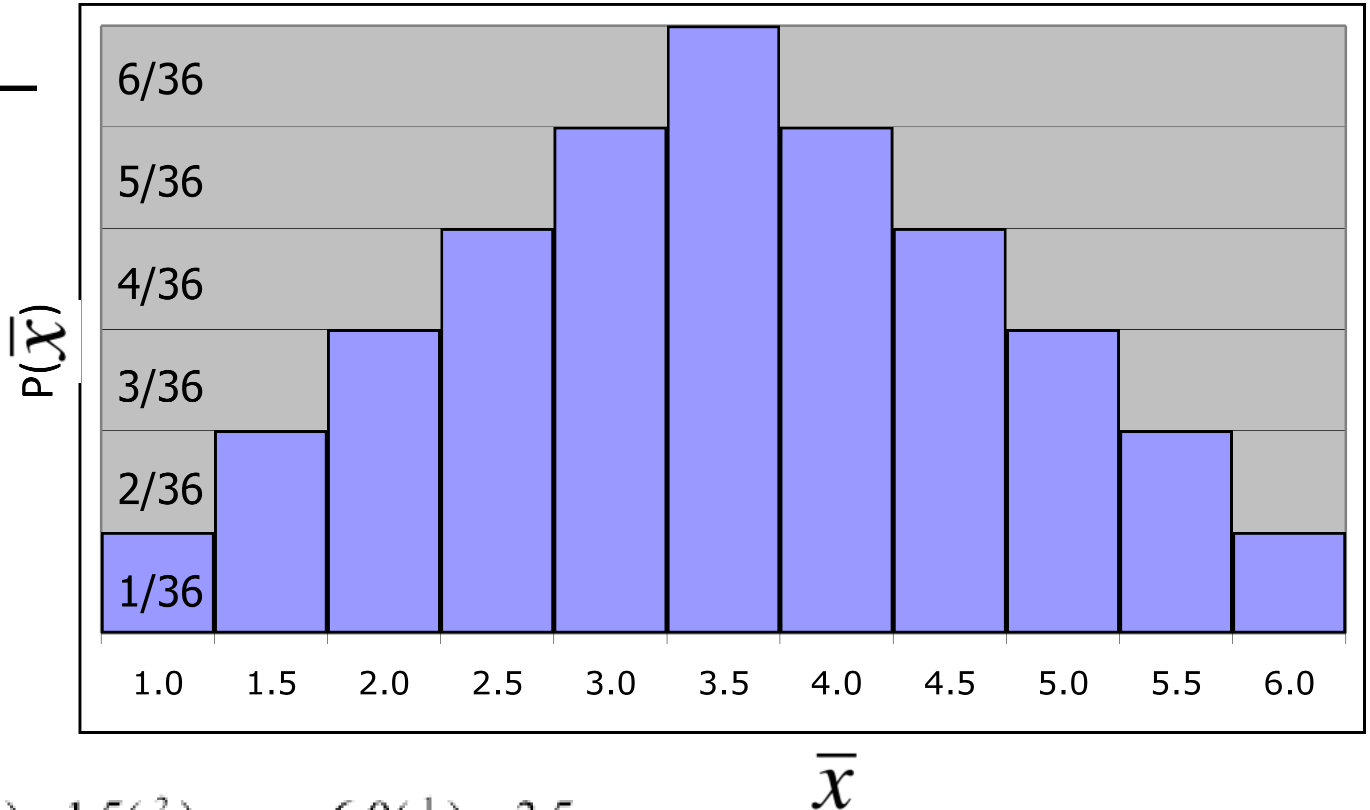
- A sampling distribution is created by looking at all samples of size n=2 (i.e. two dice) and their means…

Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

- While there are 36 possible samples of size 2, there are only 11 values for , and some (e.g. =3.5) occur more frequently than others (e.g. =1).

- The ***sampling distribution*** of $\bar{x}$ is shown below:

$\bar{x}$	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

$$\sigma^2_{\bar{x}} = \sum (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1.0 - 3.5)^2(\tfrac{1}{36}) + \ldots + (6.0 - 3.5)^2(\tfrac{1}{36}) = 1.46$$

$$\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}} = \sqrt{1.46} = 1.21$$

$n = 5$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .5833\,(= \frac{\sigma^2_x}{5})$

$n = 10$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .2917\,(= \frac{\sigma^2_x}{10})$

$n = 25$

$\mu_{\bar{x}} = 3.5$

$\sigma^2_{\bar{x}} = .1167\,(= \frac{\sigma^2_x}{25})$

Notice that $\sigma^2_{\bar{x}}$ is smaller than s_x^2. The larger the sample size the smaller $\sigma^2_{\bar{x}}$. Therefore, $\bar{X}$ tends to fall closer to m, as the sample size increases.

Generalize…

- We can generalize the mean and variance of the sampling of two dice:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \sigma^2 / 2$$

- …to **n**-dice:

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

The standard deviation of the sampling distribution is called the ***standard error***:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

LAW OF LARGE NUMBERS AND CENTRAL LIMIT THEOREM

Both are asymptotic results about the sample mean:

- Law of Large Numbers (LLN) says that as n →∞,the sample mean converges to the population mean, i.e.,

$$\text{as n} \to \infty, \bar{X} - \mu \to 0$$

- Central Limit Theorem (CLT) says that as n →∞, also the distribution converges to Normal, i.e.,

$$\text{as n} \to \infty, \ \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad \text{converges to N(0,1)}$$

- **If a population is normally distributed** with mean **μ** and standard deviation **σ**, the sampling distribution of $\overline{X}$ is also normally distributed with

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- Z-value for the sampling distribution of $\overline{X}$ is calculated:

$$Z = \frac{(\overline{X} - \mu_{\overline{X}})}{\sigma_{\overline{X}}} = \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:

$\overline{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

STUDENT’S t-DISTRIBUTION

Consider a random sample X1, X2, ..., Xn drawn from N(μ,σ2).

It is known that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ exactly distributed as N(0,1).

$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is NOT distributed as N(0,1).

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}=\frac{(\bar{X}-\mu)/\left(\sigma/\sqrt{n}\right)}{\sqrt{S^2/\sigma^2}}=\frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}=t_{n-1}$$

A different distribution for each ν= n-1 degrees of freedom (d.f.).

In statistical inference, Student’s t distribution is very important.

DISTRIBUTION OF SAMPLE VARIANCE

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Sample estimate of population variance (unbiased).

Case If Z ~ N(0,1), then $Z^2 \sim \chi_1^2$

$$\chi^2_{(n-1)} = \frac{(n-1)s^2}{\sigma^2}$$

Multiply variance estimate by n-1 to get sum of squares. Divide by population variance to normalize. Result is a random variable distributed as chi-square with (n-1) *df*.

We can use info about the sampling distribution of the variance estimate to find confidence intervals and conduct statistical tests.

F-DISTRIBUTION

Consider two independent random samples:

$X_1, X_2, \ldots, X_{n_1}$ from $N(\mu_1, \sigma_1^2)$, $Y_1, Y_2, \ldots, Y_{n_2}$ from $N(\mu_2, \sigma_2^2)$.

Then

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2} \Big/ (n_1-1)}{\dfrac{(n_2-1)S_2^2}{\sigma_2^2} \Big/ (n_2-1)}$$

has an F-distribution with n1-1 d.f. in the numerator and n2-1 d.f. in the denominator.

- F is the ratio of two independent χ2's divided by their respective d.f.'s
- Used to compare sample variances.

SAMPLING DISTRIBUTION OF A PROPORTION

- The parameter of interest for nominal data is the **proportion of times** a particular outcome (success) occurs.
- To estimate the population proportion 'p' we use the sample proportion.

The number of successes

The estimate of p = $\hat{p} = \frac{X}{n}$

- Since X is binomial, probabilities about $\hat{p}$ can be calculated from the binomial distribution.
- Yet, for inference about $\hat{p}$ we prefer to use normal approximation to the binomial whenever this approximation is appropriate.
- From the laws of expected value and variance, it can be shown that E($\hat{p}$) = p and V($\hat{p}$)=p(1-p)/n
- If both np ≥ 5 and n(1-p) ≥ 5, then

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- Z is approximately standard normally distributed.