Introduction to inference

Population vs. Sample

- Population is set of entities with some common feature.
- Sample is a representative subset of the population. It must be chosen according to:
  - Unbiasedness: same probability
  - Representativeness: same proportion
  - Dimension
- Random sample: All elements of the population have the same probability of being chosen.

Population Probability Distribution

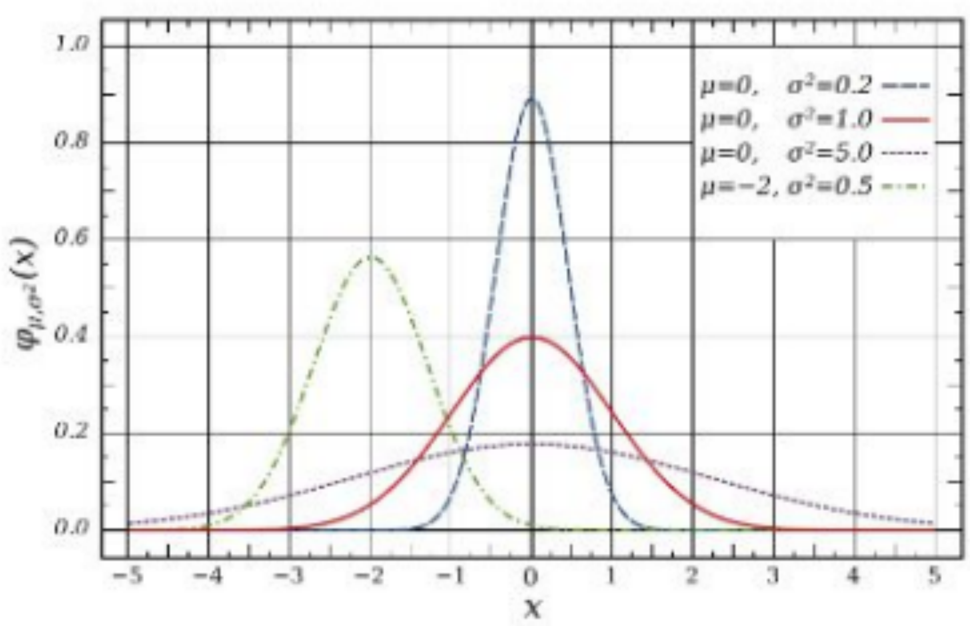
- Population probability distribution is the list of values and corresponding probabilities that a population can have.

x	**abs. freq.**	**Pr(x)**
5	1	1/5
7	2	2/5
8	1	1/5
12	1	1/5
	N=5	$\sum = 1$

- From the population probability distribution we can compute parameters, such as mean μ and standard deviation σ.

Normal Distribution

- The normal distribution is a continuous probability distribution with two parameters: μ and σ.



- The standard normal population can be used to compute the probability of a given interval for any μ and σ.

Standard Normal Distribution

- The standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$.

- Normal distributions can be **transformed** to the standard normal distribution by the formula:

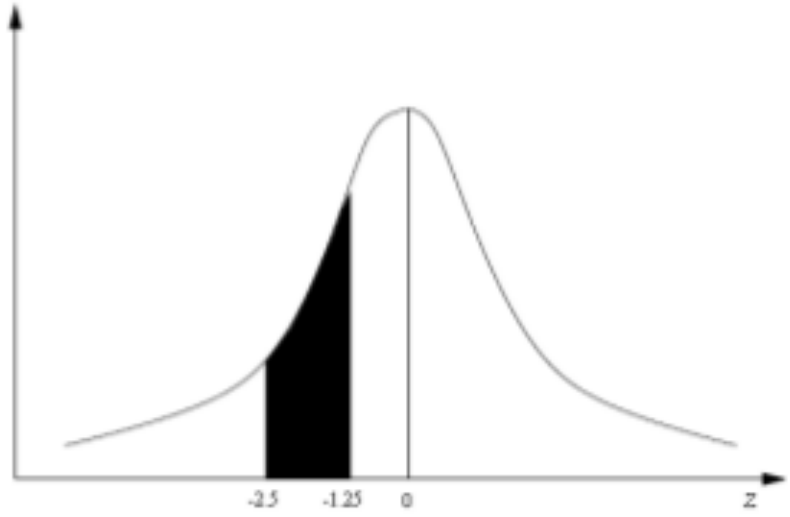
$$Z = \frac{x - \mu}{\sigma}$$

where x is the score from the original distribution

Standard Normal Distribution

- The probabilities associated to the standard normal distribution are tabulated. (more about this later on)

$\Pr(4 \leq X \leq 5) = ?$



$\Pr(-2.5 \leq Z \leq -1.25) = 9.94\%$

Sampling distributions

- A sample $x_1, \ldots, x_n$ is a representative subset of the population.

- Each element x_i is a random variable. Thus, each x_i has the same probability distribution of the population.

- The sample mean $\bar{x}$ changes according to the sample.

- Then, $\bar{x}$ is also a random variable and it has a probability distribution.

Sampling distribution of means

- Consider all samples of 3 elements (there are $\binom{5}{3}$= 10 possible samples) and compute the sample mean for each one.

$\bar{x}$	**abs. freq.**	**Pr($\bar{x}$)**
6.(3)	1	1/10
6.(6)	2	2/10
7.(3)	1	1/10
8	2	2/10
8.(3)	1	1/10
8.(6)	1	1/10
9	2	2/10
	N=10	$\sum = 1$

Sampling distribution of means

- Mean of the sampling distribution of means ($\mu_{\bar{x}}$) is equal to μ:

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

- Standard deviation of the sampling distribution of means ($\sigma_{\bar{x}}$) is equal to σ, divided by the root square of sample size (n):

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

Note: there is a correction of $\sigma_{\bar{x}}$ for large samples

Sampling distribution of means

- If a given population follows **a normal distribution** with mean μ and standard deviation σ, then the sampling distribution of means also follows a normal distribution with the following parameters:

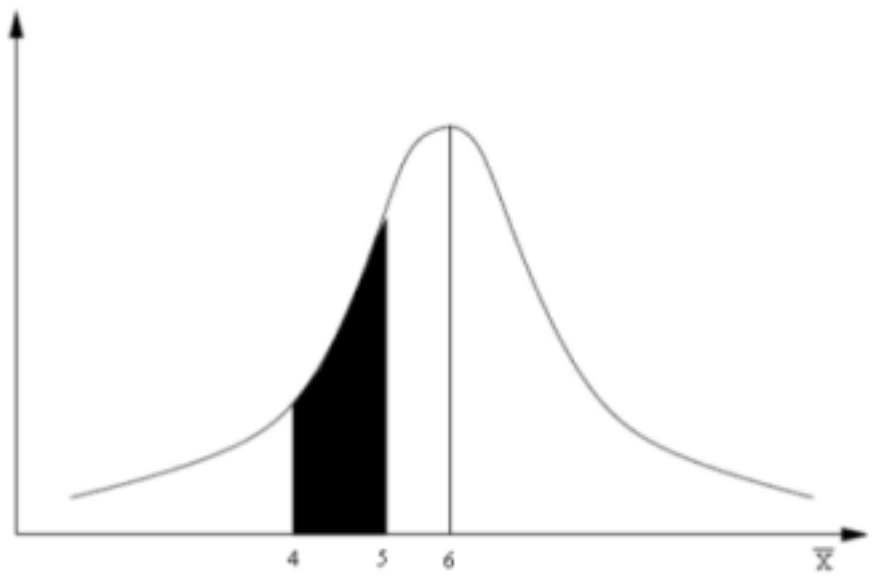
$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling distribution of means

Example: The time of user connection to my blog follows a normal distribution with a mean of 6 minutes and a standard deviation of 4 minutes. In a random sample of 25 user connections, which is the probability that they take between 4 and 5 minutes, in average?

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{5}$$



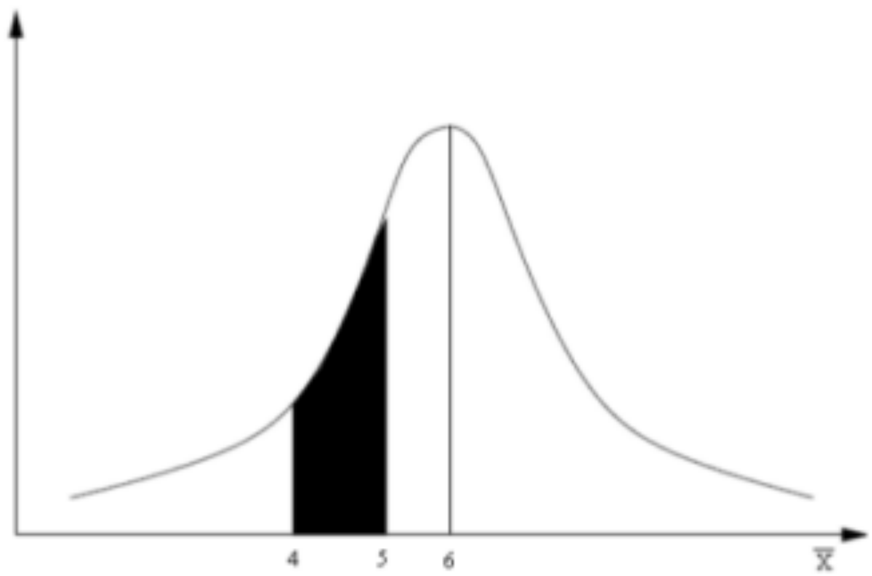
$$\Pr(4 \le \bar{x} \le 5) = ?$$

Sampling distribution of means

Example: The time of user connection to my blog follows a normal distribution with a mean of 6 minutes and a standard deviation of 4 minutes. In a random sample of 25 user connections, which is the probability that they take between 4 and 5 minutes, in average?

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{5}$$



$$\Pr(-2.5 \leq Z \leq -1.25) = 9.94\%$$

Sampling distribution of means

- If a given population with **unknown distribution** with mean μ and standard deviation σ, then the sampling distribution of means, for **increasing *n***, also follows a normal distribution with the following parameters:

$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Also known as the Central Limit Theorem

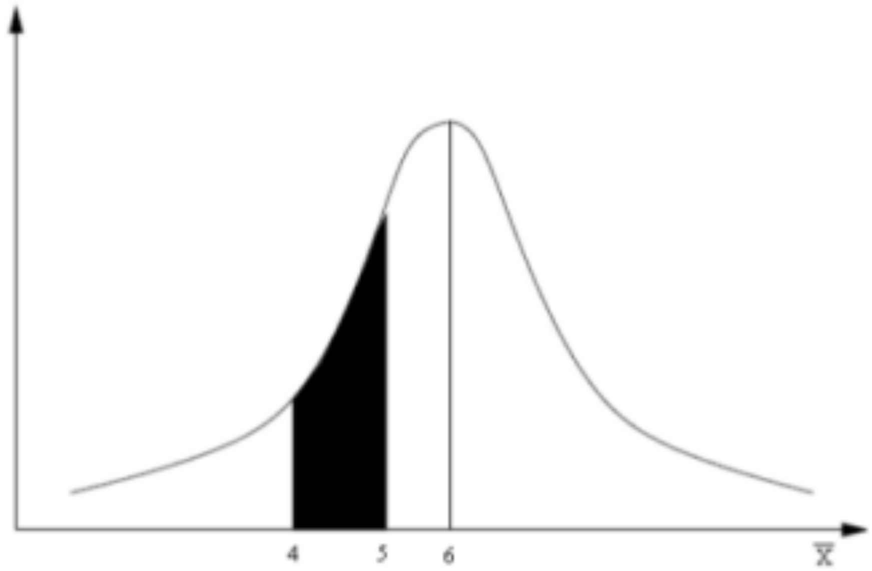
Sampling distribution of means

Example: The time of user connection to my blog follows an unknown distribution with a mean of 6 minutes and a standard deviation of 4 minutes. In a random sample of **36** user connections, which is the probability that they take between 4 and 5 minutes, in average?

$$\mu_{\bar{x}} = \mu = 6$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{6}$$

$$n > 30$$



$$\Pr\left(4 \le \bar{x} \le 5\right) = 6.55\%$$