Parameters of a distribution

- Distributions are defined by parameters. For example, the Gaussian or normal distribution is defined by the mean μ and the standard deviation σ or variance σ^2.
- A discrete distribution like the Binomial distribution is defined by parameters p and n. Example, the number of times Head occurs when a coin is tossed n times. p is the probability of a Head in a single toss.
- Populations can consist of mixtures. For example, a mixture of Gaussians. For example, if we look at weight distribution in the population we expect a mixture of two Gaussians - one for males and another for females.
- One major question in statistical inference is how to infer parameter values of the population given the values for a sample.

Measures of a distribution I

- Mean. For a discrete distribution $\mu = \sum_{i=1}^{n} x_i p_i$. For a continuous distribution: $\mu = \int_{-\infty}^{\infty} xp(x)dx$, where $p(x)$ is the pdf of X. μ is also called the expected value and often written as $E(X)$.
- $E(.)$ operator is linear. $E(aX + bY) = aE(X) + bE(Y)$.
- Median. The middle value of a distribution. For a discrete distribution assuming the data values are ordered and there are n values then:

  $$Median = \begin{cases} x_{\lceil \frac{n}{2} \rceil} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

  For continuous X the median is the value of X for which cdf is 0.5.

Measures of a distribution II

- Mode. Most frequent value for a discrete distribution. For a continuous distribution it is the value for which the pdf has the maximum value. More often when there are multiple local maxima it is referred to as a multi-modal distribution.
- Variance. $var(X) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$. For a continuous distribution $var(X) = \int_{-\infty}^{\infty}(x - \mu)^2 p(x)dx$.
- Standard deviation. $\sigma(X) = \sqrt{var(X)}$. The variance (or standard deviation) describes the spread of values in a distribution.

Population, samples, sampling distribution I

- Given the distribution type and the parameters of that distribution type for a population completely defines a population. In addition there are different measures of a population like mean, std-deviation, median, mode etc. some of which could also be parameters. For example, the mean and variance are parameters of a normal distribution.
- A **statistic** describes a measure of a sample. The sample is a subset of the population - normally chosen randomly.
- Consider picking many random samples from a population. Then any statistic of these samples is itself a RV and will have a distribution associated with it. This is called the sampling distribution. The sampling distribution of the statistic will itself have measures, in particular the mean and standard deviation (called standard error).

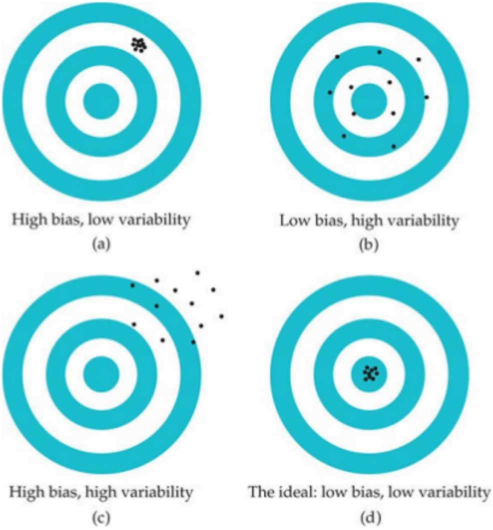
Population, samples, sampling distribution II

- The sampling distribution depends on a) the statistic being measured b) the distribution of the underlying population c) the sampling method to collect samples d) size of the sample. In practice experiments are done only on a single sample.
- Statistical inference infers unknown population parameters using sample statistics.
- Sampling distributions are at the heart of hypothesis testing and statistical inference.
- A statistic is **unbiased** if the mean of the statistic for the sampling distribution is the true value of the parameter for the population. The variability (variance is one measure of variability) or dispersion of a statistic is the spread of the statistic in the sampling distribution. There are several other measures of variability - range, std deviation, inter-quartile range, etc.

Population, samples, sampling distribution III

- Ideally we want low (or no) bias and low variability. (see fig. on later slide).
- To reduce bias in a sample we use random sampling. To reduce variance of a statistic of the sample distribution we increase the size of the sample.
- Variance of a statistic does not depend on the size of the population provided it is large enough (typically greater than 100 times sample size).
- Under-coverage and non-response can bias a sample even when it has been obtained by randomization.

Bias, variability schematic



High bias, low variability
(a)

Low bias, high variability
(b)

High bias, high variability
(c)

The ideal: low bias, low variability
(d)