Basic Statistics and Probability

Chapter 6: Sampling Distributions

- Parameters and Statistics
- Sampling Distribution
- Point Estimators

Parameters and Statistics

In Chapters 4 and 5, we characterized the probability distribution of the random variable of interest by a *parametric model*.

A parametric model has mathematical expressions which are completely known and depend on some population parameters, which are usually unknown.

- For the Bernoulli(p) r.v.

  $$X = \begin{cases} 1 & \text{if a Spaniard is favourable to Autumn and Spring clock changes} \\ 0 & \text{if he/she is against} \end{cases}$$

  the parameter p is the proportion of Spaniards in favour of the clock changes.
- We may assume that the insulin level of a healthy person follows a normal distribution with parameters μ and σ.

We can use the sample information to *make inferences* about the unknown parameters. That is what Parametric Inference is about.

A sample statistic is a value constructed using the sample. The sample statistic is frequently a numerical descriptive measure.

Many sample statistics are defined to estimate unknown population parameters:

Population parameters and corresponding statistics

	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Bernoulli proportion	p	$\bar{x}$

Parameter is a population quantity.

Statistic is a sample quantity.

Sampling Distribution

Statistics as random variables:

- The value of a sample statistic depends on the sample values $X_1, X_2, \ldots, X_n$ and is given by a formula.

  Example 6.1: BU wrist circumference

  $X \sim N(\mu, \sigma)$ is the dominant wrist circumference (in cm) of a Boston University student. To approximate the values of the parameters μ and σ, I intend to sample 4 students:

  $$X_1, X_2, X_3, X_4 \longrightarrow \bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$$

- The sample statistic has a distribution that depends on the distribution of the X_i's.

  Example 6.1 BU wrist circumference

  What is the distribution of $\bar{X}$, the mean wrist circumference in the sample of 4 students?

The distribution of the sample statistic is called its sampling distribution.

Example 6.1 BU wrist circumference

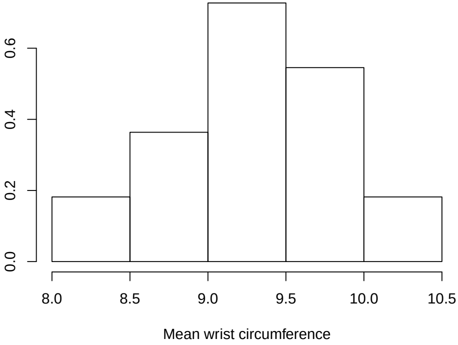
Imagine that I sample 4 students from BU repeatedly:

$$x_1 = 8.2,\ x_2 = 7.6,\ x_3 = 8.9,\ x_4 = 7.5 \rightarrow \bar{x} = 8.85$$
$$x_1 = 8.4,\ x_2 = 11.5,\ x_3 = 9.2,\ x_4 = 10.0 \rightarrow \bar{x} = 9.26$$
$$x_1 = 10.3,\ x_2 = 7.1,\ x_3 = 7.2,\ x_4 = 9.0 \rightarrow \bar{x} = 9.8$$
$$x_1 = 8.1,\ x_2 = 9.0,\ x_3 = 11.6,\ x_4 = 7.7 \rightarrow \bar{x} = 10.35$$
$$x_1 = 7.2,\ x_2 = 8.4,\ x_3 = 9.8,\ x_4 = 10.1 \rightarrow \bar{x} = 9.3$$
$$x_1 = 10.0,\ x_2 = 8.9,\ x_3 = 8.1,\ x_4 = 12.6 \rightarrow \bar{x} = 8.7$$
$$\vdots$$

and several more samples of size 4.

Example 6.1 BU wrist circumference

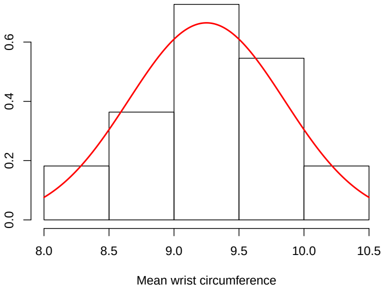
And I plot the histogram of the $\bar{x}$ values



This behaviour of $\bar{x}$ in repeated sampling is informative about the sampling distribution of $\bar{X}$.

Example 6.1 BU wrist circumference

Could the red line represent the probability density of $\bar{X}$?



Knowing the distribution of the sample statistic allows to determine the probability that this statistic takes certain values.

Example 6.1 BU wrist circumference

What is the probability that the mean dominant wrist circumference in a sample of 4 BU students is between 9 and 10 cm? $P\{9 < \bar{X} < 10\}$

Sampling distributions depend *on the statistic* and *on the population*.

Sampling distribution of the sample mean for a normal population

Let $X \sim N(\mu, \sigma)$.

For a random sample $X_1, X_2, \ldots, X_n$ of independent copies of X, the sample mean is

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The sampling distribution of $\bar{X}$ is $N(\mu, \sigma/\sqrt{n})$.

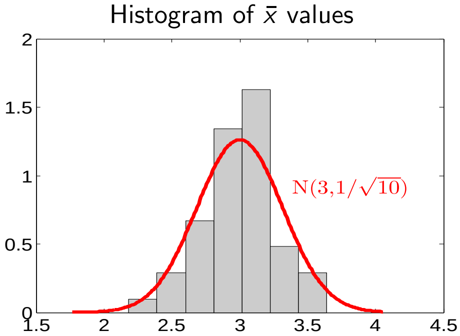
Observe that the larger n is the more concentrated the sample mean $\bar{X}$ is around the population mean μ:

$$V(\bar{X}) = \frac{\sigma^2}{n}.$$

Example 6.2:

I take 50 samples each one of size $n = 10$ from a N(3,1) distribution. I compute the corresponding 50 sample means, obtaining

3.17	3.24	3.11	3.16	2.60	2.99	2.61	2.72	2.86	3.58
2.18	3.38	2.48	3.47	3.07	2.62	3.13	2.83	3.20	3.12
3.17	3.21	2.80	2.93	2.97	3.06	2.81	2.78	2.97	3.21
3.27	2.86	2.89	2.95	2.72	3.64	3.05	3.28	2.64	2.91
3.08	3.10	2.49	3.31	2.85	3.16	3.22	2.89	3.00	3.17

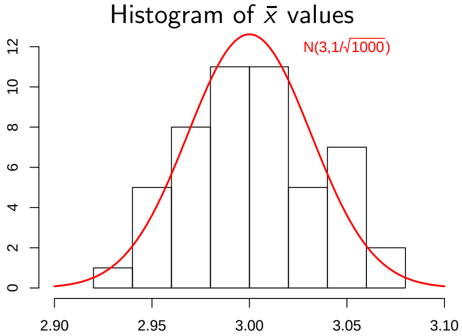


Mean of $\bar{x}$ values = 2.99 Standard deviation of $\bar{x}$ values = 0.28

Example 6.2

I take 50 samples each one of size $n = 1000$ from a N(3,1) distribution. I compute the corresponding 50 sample means, obtaining

2.98	2.98	3.04	3.01	2.96	3.00	3.01	3.05	2.94	2.94
2.99	2.95	2.94	3.06	3.02	2.99	3.02	3.03	2.98	2.98
2.96	3.01	2.99	3.07	2.96	3.02	2.98	3.07	3.05	2.98
3.02	3.00	3.05	3.00	3.01	3.01	2.99	3.04	2.98	2.96
3.00	3.04	2.99	2.97	2.99	3.01	3.03	3.02	2.96	3.01



Mean of $\bar{x}$ values = 3.00   Standard deviation of $\bar{x}$ values = 0.03

Point Estimators

A point estimator of a population parameter is a statistic (that is, a function of the sample) that estimates (approximates) the value of the parameter.

The difference between the point estimator and the target parameter is the error of estimation.

Some usual population parameters and point estimators are:

	Parameter	Point estimator	Error
Mean	μ	$\hat{\mu} = \bar{x}$	$\bar{x} - \mu$
Variance	σ^2	$\hat{\sigma}^2 = s^2$	$s^2 - \sigma^2$
Standard deviation	σ	$\hat{\sigma} = s$	$s - \sigma$
Binomial proportion	p	$\hat{p} = \bar{x}$	$\bar{x} - p$

We may guess how large the error of estimation is likely to be by looking at the sampling distribution of the statistic.

As point estimators are sampling statistics, they have a sampling distribution.

To compare two point estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, used to estimate the same parameter θ, we should look at the sampling distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ and compare their properties.

1. Bias. If the expected value of the parameter estimator $\hat{\theta}$ coincides with the parameter θ (the estimation target), then the estimator $\hat{\theta}$ is unbiased:

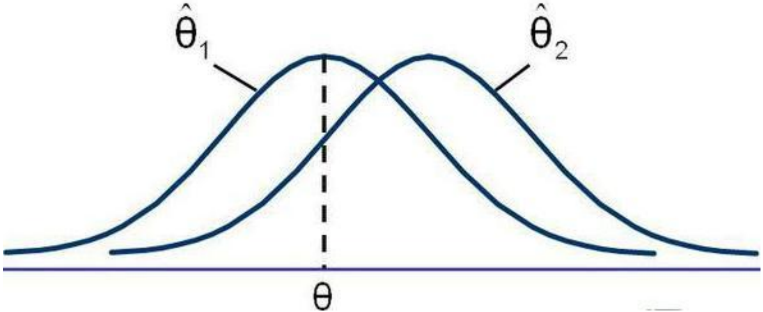
$$E(\hat{\theta}) = \theta.$$

If the expected value of the estimator $\hat{\theta}$ is not equal to the parameter θ, then the estimator $\hat{\theta}$ is biased:

$$E(\hat{\theta}) \neq \theta.$$

Unbiased estimators are generally preferred over biased ones.

$\hat{\theta}_1$ is an unbiased estimator, $\hat{\theta}_2$ is biased:

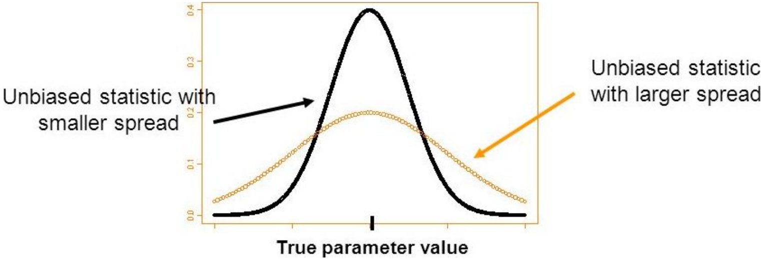


The blue curves represent the probability densities of $\hat{\theta}_1$ and $\hat{\theta}_2$.

2. Dispersion. The standard error (s.e.) of an estimator $\hat{\theta}$ is the standard deviation of the distribution of $\hat{\theta}$ and measures the spread of $\hat{\theta}$. If two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of θ are both unbiased

$$E(\hat{\theta}_1) = \theta \qquad \text{and} \qquad E(\hat{\theta}_2) = \theta,$$

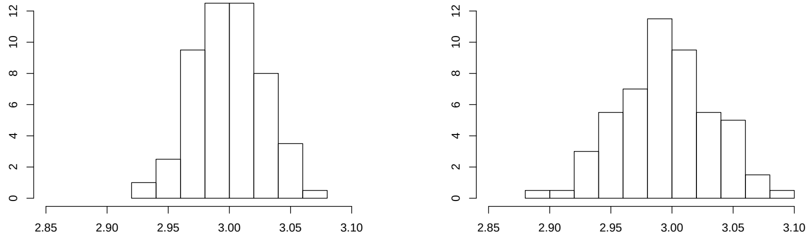
then we choose the estimator with the lowest standard deviation.



If $V(\hat{\theta}_1) < V(\hat{\theta}_2)$, then the probability that $\hat{\theta}_1$ is close to θ is higher than the probability that $\hat{\theta}_2$ is close to θ.

Example 6.3:

We take 100 samples each one of size $n = 1000$ from a $N(3, 1)$. For each sample we compute the sample mean $\bar{x}$ and the median M and we plot the histograms of the 100 means (left) and medians (right). Which is a better estimator of $\mu = 3$, the mean or the median?



Sample mean of the 100 $\bar{x}$'s = 3.00 Sample mean of the 100 M's = 3.0

s.d. of the 100 $\bar{x}$'s = 0.03 s.d. of the 100 M's = 0.04

Sampling Distribution of $\bar{X}$. Central Limit Theorem

We said that, if $X \sim N(\mu, \sigma)$, then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

of a sample of size n from X follows a $N(\mu, \sigma/\sqrt{n})$.

Even if X is not Gaussian this is approximately true for a sufficiently large sample size n $(n \geq 30)$.

Central Limit Theorem

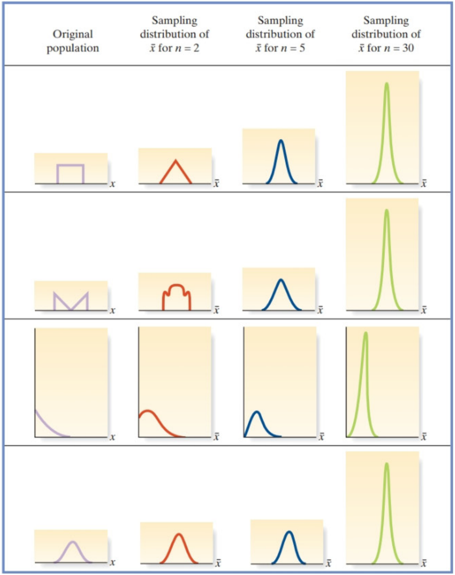
Let X be **any** r.v. with mean μ and standard deviation σ.

Let $\bar{X}$ be the the sample mean of n independent copies of X.

For n "large enough" the sampling distribution of $\bar{X}$ is approximately

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The larger the sample size n, the better will be the normal approximation to the distribution of $\bar{X}$.



Example 6.4: Serum Cholesterol

According to the National Center for Health Statistics, the distribution of serum cholesterol levels for 20- to 74-year-old males living in the United States has mean 211 mg/dl, and a standard deviation of 46 mg/dl. We are planning to collect a sample of 25 individuals and measure their cholesterol levels. What is the probability that our sample average will be above 230?