Sampling Distributions

Up to now we have assumed that we knew all we needed to know about the probability distribution of a random variable. Using that knowledge we computed the mean, variance, standard deviation, and probabilities associated with the random variable.

For example, in the case of a binomial random variable with $n = 10$ and $p = 0.30$ we have the following:

Mean: $\mu = np =$ _______; Variance: $\sigma^2 = np(1-p) =$ _______;

Standard Deviation: $\sigma = \sqrt{npq} =$ _______; and $P(x \le 5) \approx$ _______.

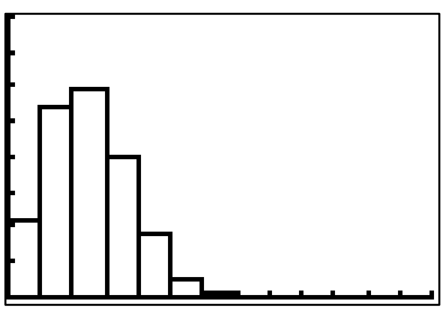
In the case of a normal distribution with $\mu = 0$ and $\sigma = 1$ we have the following:

$P(z < 0.67) \approx$ ________.

In real applications it is unlikely that the true population mean and standard deviation are known. Hence they must be estimated from samples. The numerical values that describe probability distributions are called ***parameters***. So the numbers p, μ, and σ above are examples of parameters,

We have also discussed the sample mean $\bar{x}$, sample variance s^2, and sample standard deviation s, which are descriptive measures calculated from a sample. We will frequently use those ***sample statistics*** to make inferences about the parameters of a population.

Example 1: Consider the experiment that consists of rolling a fair die twelve times. We will call the result of the experiment a success if one dot appears on the top face of the die. The random variable x represents the number of successes in twelve rolls of the die. So, we have a binomial probability distribution with $p = 1/6$ and $n = 12$. A graph of this distribution is shown below.



We know the mean, or expected, value of x is μ = ______,

and the standard deviation for the distribution is σ = _____.

However, μ and σ are theoretical values based on the assumptions stated in the example. Suppose we wanted to estimate the values of μ and σ based on statistics calculated from a sample?

Suppose we create a sample by conducting the experiment five times and the number of successes on successive experiments is: 3, 2, 2, 0, 1. The sample mean $\bar{x}$ is 1.6.

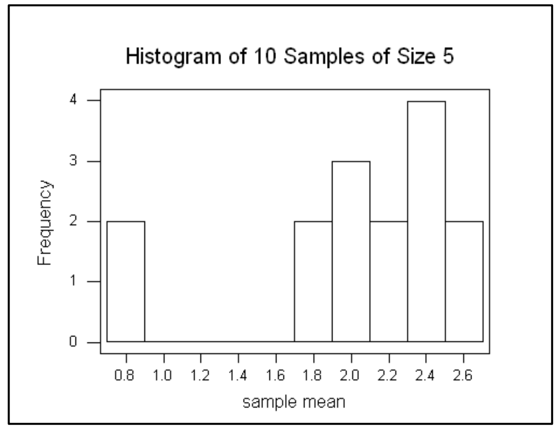
Now suppose we create a second sample of the same size and the number of successes on successive experiments is: 1, 5, 3, 4, 4. The sample mean $\bar{x}$ in this case is 3.4.

It is important to note that different samples from the same population of measurements can yield very different sample statistics. So, sample statistics, like $\bar{x}$, s^2, as s are themselves random variables. If our sampling experiment is repeated a very large number of times, the resulting histogram of sample means would be the approximate probability distribution of $\bar{x}$. If $\bar{x}$ is a good estimator of μ, we expect the values of $\bar{x}$ to cluster around μ. This probability distribution is called a ***sampling distribution***.

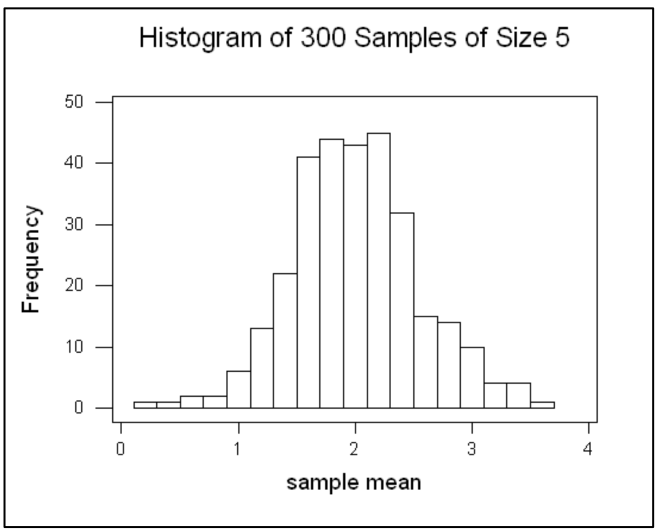
Here is the result of a MINITAB simulation of 10 samples of size five. Each sample consists of doing the experiment five times.

Worksheet 1 ***

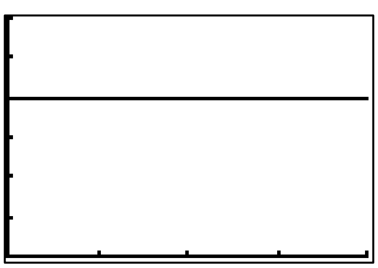
	C1	C2	C3	C4	C5	C6
↓	experiment 1	experiment 2	experiment 3	experiment 4	experiment 5	sample mean
1	2	4	3	4	0	2.0
2	1	3	1	0	1	1.8
3	3	2	3	1	1	2.4
4	1	2	1	0	1	2.4
5	1	0	1	2	2	0.8
6	0	3	2	3	1	2.2
7	4	1	3	0	2	2.6
8	1	3	4	2	3	2.4
9	3	0	3	2	4	1.8
10	2	2	1	1	2	0.8
11						2.6



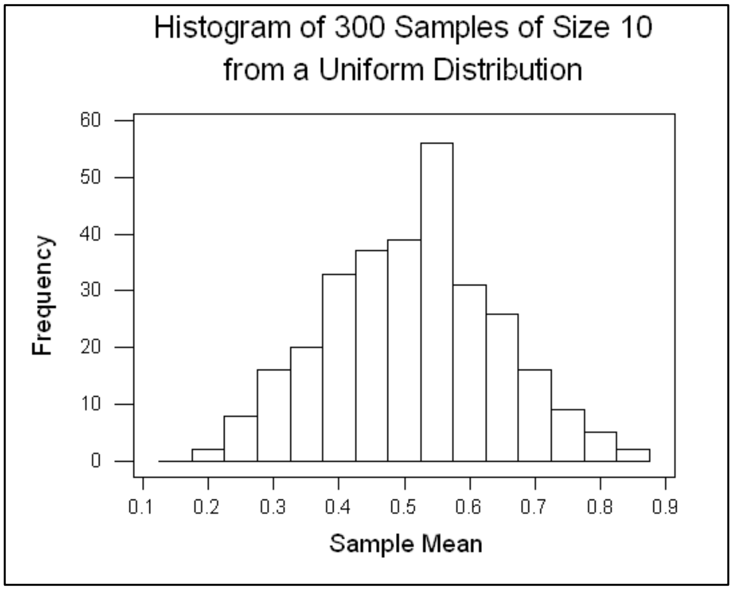
Below we look at a frequency histogram for a sampling distribution resulting from a MINITAB simulation of 300 samples of size five. The mean for that derived sampling distribution is $\bar{x} \approx$ 2.0133.



Example 2. Suppose we select a number at random from the interval 0 < x < 1. This distribution, known as a ***uniform distribution*** and its probability distribution f(x) is shown below.



We will use MINTAB to generate 300 samples from the population, each with 10 observations. We will then compute $\bar{x}$ and the median M for each of the 1000 samples. We will then approximate $\bar{x}$ and M for the sampling distributions of those statistics and compare those values to the population mean μ which is 0.5.



The histogram of the distribution of sample means is shown to the left. The mean of the means of the 300 samples is approximately 0.51177 which provides a fairly good estimate of the population mean μ which is 0.5

(See pp. 252-253 for instructions on how to perform this simulation.)

Example 3. Suppose you were given a die that was claimed to be fair. You proceeded to perform the following experiment five times. Roll the die twelve times and record the number times one dot comes up on the top face. The mean number of times one dot appeared on the top face in your five trials was one. Should you be surprised? How strongly would you believe the claim that the die is fair?