

VII. Sampling Distributions

A. Some Important Definitions (reminders)

1. Population - Collection of all possible elements of interest
2. Census - Collection of the values for all variables of interest that correspond to all elements of a population. The size of the census (number of elements in the population) is usually denoted N
3. Parameter - A summary measure used to describe values of a variable (i.e., a characteristic) for the entire population

4. Sample - a collection of elements that comprise a subset of the population. The size of the sample (number of elements to be included) is usually denoted n
5. Statistical Inference - using data obtained through sampling to estimate the value of or test a hypothesis about a parameter (i.e., use of inductive logic)
6. Sampling With Replacement - selection of sample objects where a sample object is returned to the population after selection (and could possibly be chosen again)
7. Sampling Without Replacement - selection of sample objects where a sample object is not returned to the population after selection (and cannot possibly be chosen again)

B. Methods of Sampling

Simple Random Sampling - each possible sample of size n chosen from a population of size N has an equal probability of being selected - this is the most common and straight-forward of all probability sampling methods.

Stratified Random Sampling - a probability sampling method for which we divide the population into homogeneous strata and take a simple random sample from each strata

Cluster Sampling - a probability sampling method for which we divide the population into heterogeneous clusters (usually by proximity) and take a census from randomly selected clusters.

Systematic Sampling - a probability sampling method for which we randomly select the first element then subsequently select every k^{th} element.

Convenience (Chunk) Sampling - a nonprobability sampling method for which elements are selected on the basis of their ease of collection.

Judgement Sampling - a nonprobability sampling method for which elements are selected on the basis of the sampler's opinion of their appropriateness.

C. Point Estimation

1. Point Estimate - a single numerical value used as an estimate of a parameter.

2. Point Estimator - the sample statistic that provides the point estimate of a parameter.

Some point estimators (and the parameters they estimate) include

<u>Parameter</u>	<u>Point Estimator</u>
μ	\bar{x}
σ	s
p	\bar{p}

3. Precision - the exactness of an estimator.

4. Accuracy - the correctness of an estimator.

What is the relationship between precision and accuracy?

Point estimators are

- Perfectly precise
- Almost certainly inaccurate

5. Desirable Characteristics of Point Estimators

Unbiasedness - the expected value of the sample (point) estimators (statistic) equals the population value (parameter)

What estimator would ever be biased?

Sample Maximum is a biased estimator of Population Maximum!

Suppose you have a population that consists of the four values 2, 5, 3, 9. The population maximum is obviously 9.0.

Now take each possible sample of 3 elements (without replacement) and find the sample maximum for each:

$$S_1 = \{2, 5, 3\}, \max = 5.0 \quad S_2 = \{2, 5, 9\}, \max = 9.0$$

$$S_3 = \{2, 3, 9\}, \max = 9.0 \quad S_4 = \{5, 3, 9\}, \max = 9.0$$

The mean of these estimates is 8.0, which does not equal the population maximum of 9.0!

Are there any other biased estimators?

Sample Median is a biased estimator of Population Median!

Suppose you have a population that consists of the five values 1, 2, 19, 20, 21. The population median is obviously 19.0.

Now take each possible sample of 4 elements (without replacement) and find the sample median for each:

$$S_1 = \{1, 2, 19, 20\}, m_d = 10.5 \quad S_2 = \{1, 2, 19, 21\}, m_d = 10.5$$

$$S_3 = \{1, 2, 20, 21\}, m_d = 11.0 \quad S_4 = \{1, 19, 20, 21\}, m_d = 19.5$$

$$S_5 = \{2, 19, 20, 21\}, m_d = 19.5$$

The mean of these estimates is 14.2, which does not equal the population median of 19.0!

So what is an example of an unbiased estimators?

Sample Mean is an unbiased estimator of Population Mean!

Suppose you have a population that consists of the five values 1, 2, 19, 20, 21. The population mean is 12.6.

Now take each possible sample of 4 elements (without replacement) and find the sample mean for each:

$$S_1 = \{1, 2, 19, 20\}, \bar{x} = 10.50 \quad S_2 = \{1, 2, 19, 21\}, \bar{x} = 10.75$$

$$S_3 = \{1, 2, 20, 21\}, \bar{x} = 11.00 \quad S_4 = \{1, 19, 20, 21\}, \bar{x} = 15.25$$

$$S_5 = \{2, 19, 20, 21\}, \bar{x} = 15.50$$

The mean of these estimates is 12.6, which does equal the population mean of 12.6!

This will always be true for any unbiased estimator (no matter what the sample size)!

Are there any other important unbiased estimators?

Sample Proportion is an unbiased estimator of Population Proportion!

Suppose you have a population that consists of the five values A, B, B, A, B. The population proportion of A is $p = 2/5 = 0.40$.

Now take each possible sample of 4 elements (without replacement) and find the sample proportion of A for each:

$S_1 = \{A, B, B, A\}, \bar{p} = 0.50$ $S_2 = \{A, B, B, B\}, \bar{p} = 0.25$

$S_3 = \{A, B, A, B\}, \bar{p} = 0.50$ $S_4 = \{A, B, A, B\}, \bar{p} = 0.50$

$S_5 = \{B, B, A, B\}, \bar{p} = 0.25$

The mean of these estimates is 0.40, which does equal the population proportion of 0.40!

Again, *this will always be true for any unbiased estimator (no matter what the sample size)!*

5. Desirable Characteristics of Point Estimators (continued)

Consistency - the probability that the value of the point estimate falls within some given range about the parameter increases to 1 as the sample size grows

6. Sampling Error - the difference between an unbiased point estimate and the actual value of the parameter being estimated

D. Sampling Distributions

1. Sampling Distribution - the probability distribution associated with a *statistic*. The most commonly used sampling distributions are those for \bar{x} and \bar{p} .

2. Standard Error - the standard deviation of a sampling distribution (i.e., of a statistic).

3. The Sampling Distribution of the Sample Mean - the probability distribution associated with all possible values of the sample mean \bar{x} .

- the expected value of this distribution is μ (i.e., $E[\bar{x}] = \mu_x = \mu$) and is sometimes denoted $\mu_{\bar{x}}$.
- the standard deviation (also called the standard error) of this distribution is

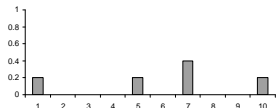
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

if the population is 'infinite' in the statistical sense, i.e.

$$\frac{n}{N} \leq 0.05$$

Example: Suppose we have a very small population of the five elements 1, 5, 7, 7, and 10:

Probability Distribution

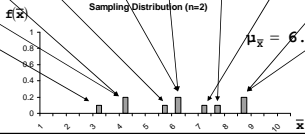


Then there are ten possible samples of size $n = 2$

$$S = \{1, 5, 7, 7, 10\} \rightarrow \mu = 6.0, \sigma = 2.966$$

$S_1 = \{1,5\}$ $S_2 = \{1,7\}$ $S_3 = \{1,7\}$ $S_4 = \{1,10\}$ $S_5 = \{5,7\}$ $S_6 = \{5,7\}$ $S_7 = \{5,10\}$ $S_8 = \{7,7\}$ $S_9 = \{7,10\}$ $S_{10} = \{7,10\}$
 $\bar{x}_1 = 3.0$ $\bar{x}_2 = 4.0$ $\bar{x}_3 = 4.0$ $\bar{x}_4 = 5.5$ $\bar{x}_5 = 6.0$ $\bar{x}_6 = 6.0$ $\bar{x}_7 = 7.5$ $\bar{x}_8 = 7.0$ $\bar{x}_9 = 8.5$ $\bar{x}_{10} = 8.5$

Sampling Distribution ($n=2$)



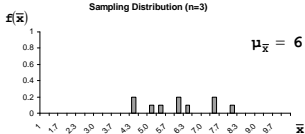
$$\mu_{\bar{x}} = 6.0, \sigma_{\bar{x}} = 2.098$$

There are also ten possible samples of size $n = 3$

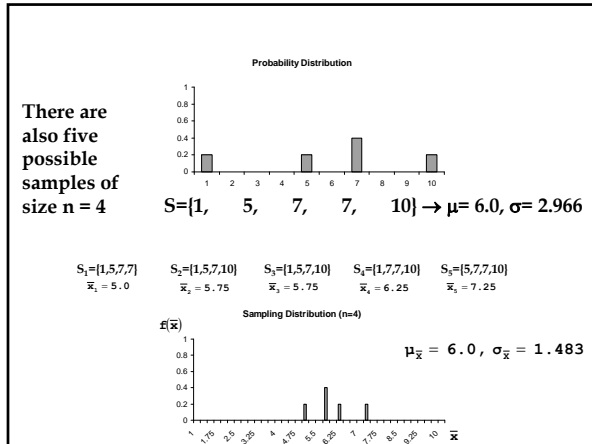
$$S = \{1, 5, 7, 7, 10\} \rightarrow \mu = 6.0, \sigma = 2.966$$

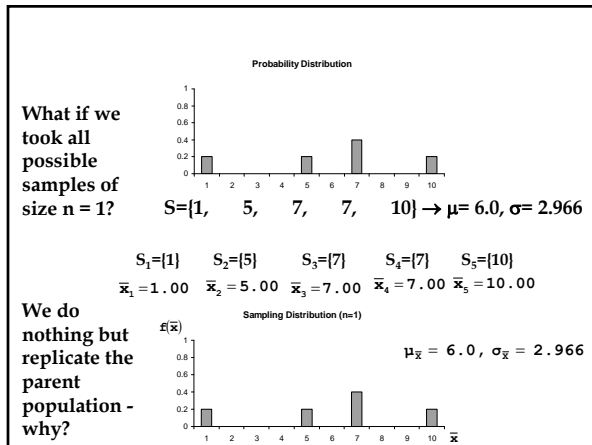
$S_1 = \{1,5,7\}$ $S_2 = \{1,5,7\}$ $S_3 = \{1,5,10\}$ $S_4 = \{1,7,7\}$ $S_5 = \{1,7,10\}$ $S_6 = \{1,7,10\}$ $S_7 = \{5,7,7\}$ $S_8 = \{5,7,10\}$ $S_9 = \{5,7,10\}$ $S_{10} = \{7,7,10\}$
 $\bar{x}_1 = 4.3$ $\bar{x}_2 = 4.3$ $\bar{x}_3 = 5.3$ $\bar{x}_4 = 5.0$ $\bar{x}_5 = 6.0$ $\bar{x}_6 = 6.0$ $\bar{x}_7 = 6.3$ $\bar{x}_8 = 7.3$ $\bar{x}_9 = 7.3$ $\bar{x}_{10} = 8.0$

Sampling Distribution ($n=3$)



$$\mu_{\bar{x}} = 6.0, \sigma_{\bar{x}} = 1.712$$



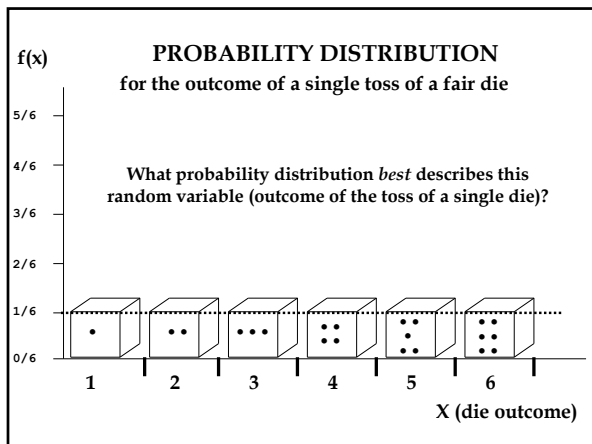


Suppose we throw a die one time. and record the value we obtain on each toss. How many potential outcomes are there?

$S = \{1, 2, 3, 4, 5, 6\}$

Since each of these outcomes is equally likely, the probability of any single outcome is

$\frac{1}{6} = 0.16666\bar{6}$



Here is a hint:

$$f(x) = \begin{cases} \frac{1}{6} & x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

It is a discrete uniform distribution!

The mean and standard deviation for this probability distribution are

$$\mu = E(x) = \sum_{x=1}^6 x f(x)$$

$$= \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = 3.5$$

and

$$\sigma^2 = \sum_{x=1}^6 (x - \mu)^2 f(x)$$

$$= \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2$$

$$+ \frac{1}{6}(4 - 3.5)^2 + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2 = 2.91667,$$

$$\sigma = \sqrt{\sigma^2} = 1.708$$

Now suppose we throw a die two times. How many potential outcomes are there?

This can be a Multiple-Step Experiments for which the number of steps is $k = 2$. Then the number of potential outcomes is:

$$(n_1)(n_2) \dots (n_k) = n_1 * n_2 = 6 * 6 = 36$$

where n_j represents the number of potential outcomes on the j^{th} trial.

Since each of these outcomes is equally likely, the probability of any single outcome is

$$\frac{1}{36} = 0.0277\bar{7}$$

Now suppose we are interested in the mean of the sides showing on our two tosses.

The mistake that most people make when calculating probabilities for such a problem is that they think of the sample space

$$\mathcal{S}_x = \mathcal{S}_{x_1+x_2} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

divide by 2 to
calculate the
sample mean

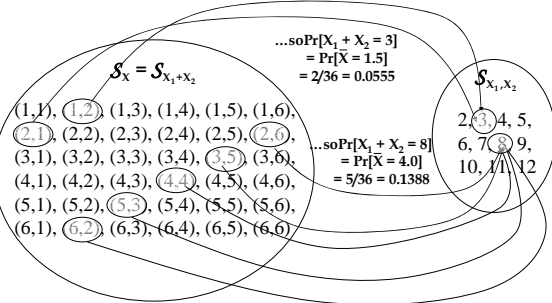
as uniformly distributed instead of considering the original sample space, which is

$$\mathcal{S}_{x_1, x_2} = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

divide the sum by 2 to
calculate the sample mean

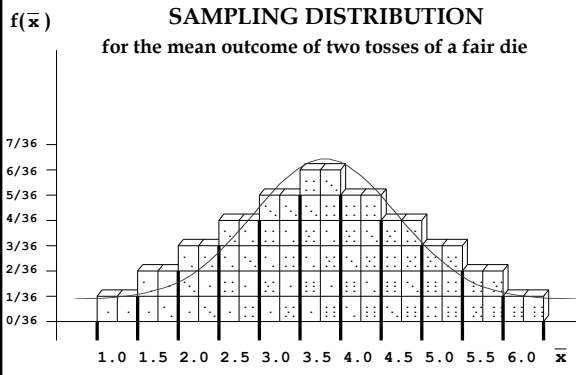
The sample space for the sum (or the mean) mean of the faces of two die could be represented graphically by

How many ways are there to roll two die at get a sum of 3 (i.e., a mean of 1.5)?



How many ways are there to roll two die at get a sum of 8 (i.e., a mean of 4.0)?

The sampling distribution for \bar{x} = the mean of the faces of two die could be represented graphically by



The probability distribution function for the mean outcome when throwing two dice looks like this:

$$f(x) = \begin{cases} 1/36 & \bar{x} = 1.0 \\ 2/36 & \bar{x} = 1.5 \\ 3/36 & \bar{x} = 2.0 \\ 4/36 & \bar{x} = 2.5 \\ 5/36 & \bar{x} = 3.0 \\ 6/36 & \bar{x} = 3.5 \\ 5/36 & \bar{x} = 4.0 \\ 4/36 & \bar{x} = 4.5 \\ 3/36 & \bar{x} = 5.0 \\ 2/36 & \bar{x} = 5.5 \\ 1/36 & \bar{x} = 6.0 \\ 0.0 & \text{otherwise} \end{cases}$$

The original population for this problem is the collection of possible values for a single thrown die. The mean and standard deviation for this distribution are $\mu=3.50$ and $\sigma=1.708$.

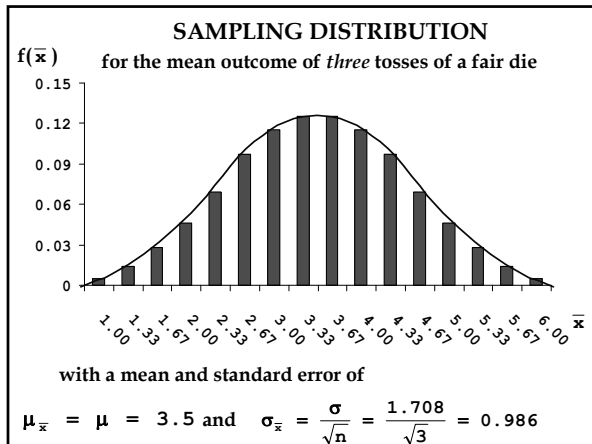
Thus the sampling distribution for the mean of the outcome when the die is thrown twice has a mean and standard deviation of

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

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{2}} = 1.208$$

But what would happen if we increase the number of tosses per trial to three?



The probability distribution function for throwing three dice looks like this:

$f(x) =$	$\frac{1}{216}$	$\bar{x} = 1.00$
	$\frac{3}{216}$	$\bar{x} = 1.33$
	$\frac{6}{216}$	$\bar{x} = 1.67$
	$\frac{10}{216}$	$\bar{x} = 2.00$
	$\frac{15}{216}$	$\bar{x} = 2.33$
	$\frac{21}{216}$	$\bar{x} = 2.67$
	$\frac{25}{216}$	$\bar{x} = 3.00$
	$\frac{27}{216}$	$\bar{x} = 3.33$
	$\frac{27}{216}$	$\bar{x} = 3.67$
	$\frac{25}{216}$	$\bar{x} = 4.00$
	$\frac{21}{216}$	$\bar{x} = 4.33$
	$\frac{15}{216}$	$\bar{x} = 4.67$
	$\frac{10}{216}$	$\bar{x} = 5.00$
	$\frac{6}{216}$	$\bar{x} = 5.33$
	$\frac{3}{216}$	$\bar{x} = 5.67$
	$\frac{1}{216}$	$\bar{x} = 6.00$
	0.0	otherwise

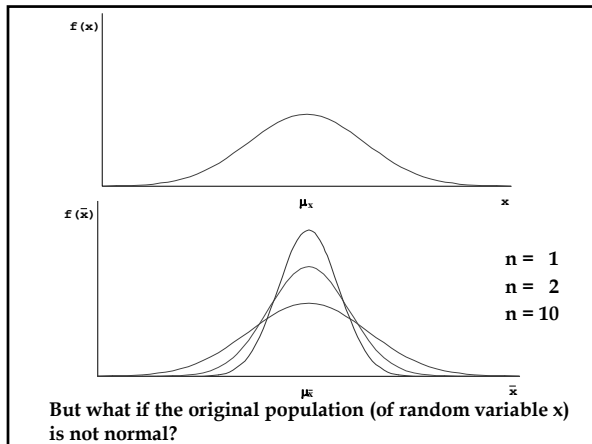
But how is this random variable (the mean of the outcomes of three thrown dice) distributed?

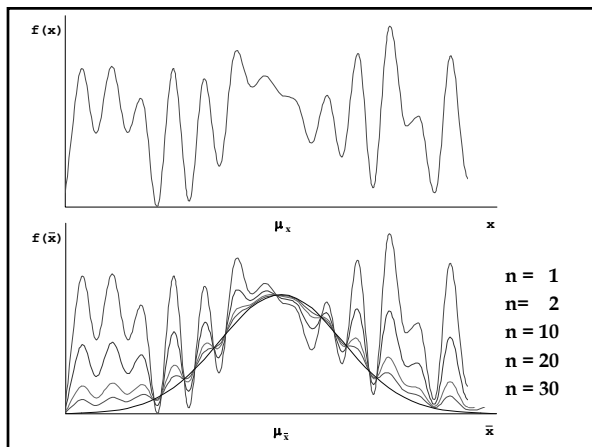
Central Limit Theorem - when selecting a simple random sample from a population, the sampling distribution of \bar{x} can be approximated by a normal probability distribution as the sample size becomes large no matter how the original population is distributed.

We can assume that any sample of at least $n = 30$ is sufficient to assure that the central limit theorem will force the potential values of \bar{x} to be normally distributed.

Why does this happen?

- If the original population (of random variable x) is normal

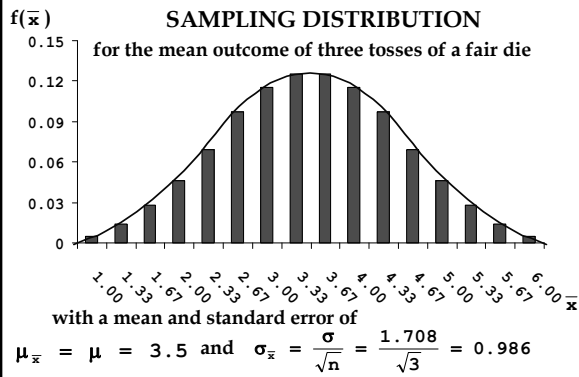




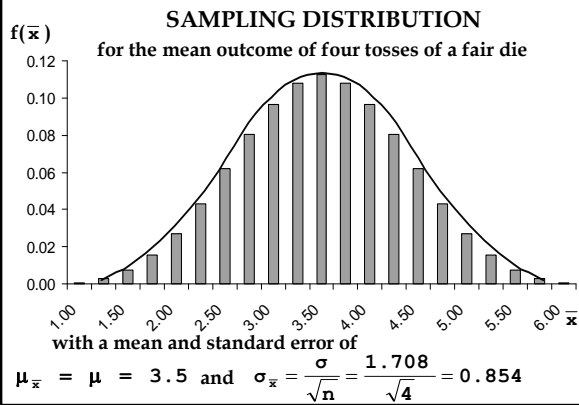
So what are the ramifications for the dice problem (or any other problem)?

- If the sample size is sufficiently large ($n \geq 30$, i.e. a single trial consists of at least 30 tosses) then we can use the normal probability distribution to describe probabilities for potential values of the sample mean!
- Even if the original (parent) population is extremely non-normal, we can still talk about the probability that the sample mean will be within a given range (if $n \geq 30$). Since we are usually much more interested in summary statistics (such as the mean) as opposed to individual observations, this result is very powerful!

Remember what happened when we simply increased the number of flips per trial to 3 and calculated sample means?



What if we increase the number of flips per trial to 4?



The probability distribution function for throwing four dice looks like this:

What is happening to the shape of the sampling distribution as we increase the number of times we throw the dice?

$\frac{1}{1296}$	$\bar{x} = 1.00$
$\frac{4}{1296}$	$\bar{x} = 1.25$
$\frac{10}{1296}$	$\bar{x} = 1.50$
$\frac{20}{1296}$	$\bar{x} = 1.75$
$\frac{35}{1296}$	$\bar{x} = 2.00$
$\frac{56}{1296}$	$\bar{x} = 2.25$
$\frac{80}{1296}$	$\bar{x} = 2.50$
$\frac{104}{1296}$	$\bar{x} = 2.75$
$\frac{125}{1296}$	$\bar{x} = 3.00$
$\frac{140}{1296}$	$\bar{x} = 3.25$
$\frac{146}{1296}$	$\bar{x} = 3.50$
$\frac{140}{1296}$	$\bar{x} = 3.75$
$\frac{125}{1296}$	$\bar{x} = 4.00$
$\frac{104}{1296}$	$\bar{x} = 4.25$
$\frac{80}{1296}$	$\bar{x} = 4.50$
$\frac{56}{1296}$	$\bar{x} = 4.75$
$\frac{35}{1296}$	$\bar{x} = 5.00$
$\frac{20}{1296}$	$\bar{x} = 5.25$
$\frac{10}{1296}$	$\bar{x} = 5.50$
$\frac{4}{1296}$	$\bar{x} = 5.75$
$\frac{1}{1296}$	$\bar{x} = 6.00$
0.0	otherwise

Example: If we toss the die 40 times, the mean and standard deviation of \bar{x} will be

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

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{40}} = 0.270$$

What is the probability that the mean value will be:

between 3.25 and 3.50?

$$\begin{aligned} P(3.25 \leq \bar{x} \leq 3.50) &= P\left(\frac{3.25 - 3.50}{0.270} \leq \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{3.50 - 3.50}{0.270}\right) \\ &\approx P(-0.93 \leq z \leq 0.00) = 0.3238 \end{aligned}$$

less than 3.25?

$$\begin{aligned} P(\bar{x} \leq 3.25) &= P(-\infty \leq \bar{x} \leq 3.25) \\ &= P\left(\frac{-\infty - 3.50}{0.270} \leq \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{3.25 - 3.50}{0.270}\right) \\ &\approx P(-\infty \leq z \leq 0.93) \\ &= 0.1762 \end{aligned}$$

between 3.25 and 3.60?

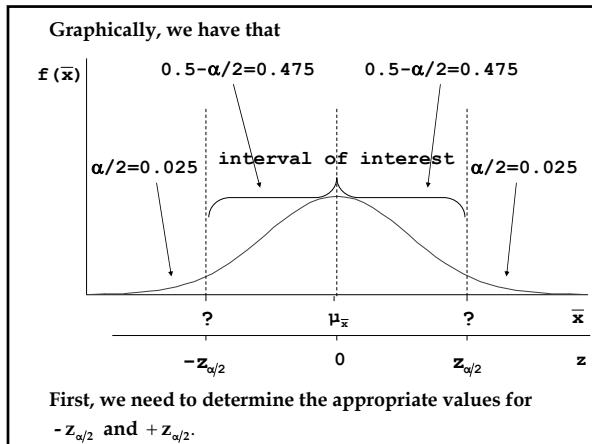
$$\begin{aligned} P(3.25 \leq \bar{x} \leq 3.60) &= P\left(\frac{3.25 - 3.50}{0.270} \leq \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{3.60 - 3.50}{0.270}\right) \\ &= P(-0.93 \leq z \leq 0.37) \\ &= P(-\infty \leq z \leq 0.37) - P(-\infty \leq z \leq -0.93) \\ &= 0.6443 - 0.1762 = 0.4681 \end{aligned}$$

Tolerance Interval - a range of values of a random variable that contains a specified proportion of a population. These are often derived so they are symmetric wrt the center of the variable's probability distribution.

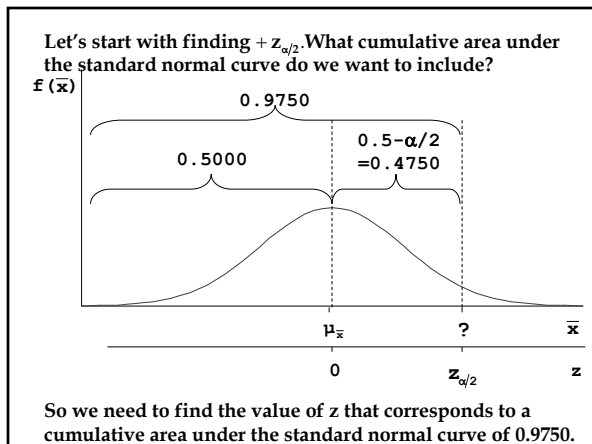
What interval, symmetric about the mean, will hold 95% of all possible values of \bar{x} ?

Let's define

- The proportion of population elements *not included* in the tolerance interval to be α (this notation will be useful later).
- The values of the standard normal variable z that produce such an interval to be $-z_{\alpha/2}$ and $+z_{\alpha/2}$.



At this point we are going to use the inverse standard normal function (i.e., we are going to use the standard normal table in reverse).
 Why? We have been using the tables to find the cumulative area under the standard normal curve for a given value of z - now we wish to find the value of z that results in a given cumulative area under the standard normal curve!



How will we find the actual values of \bar{x} that correspond to these values of z ? By inverting the z transformation!

$$z = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} \rightarrow z\sigma_{\bar{x}} = \bar{x} - \mu_x \rightarrow \mu_x + z\sigma_{\bar{x}} = \bar{x}$$

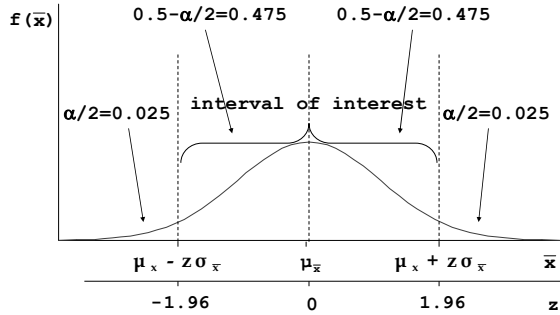
and

$$-z = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} \rightarrow -z\sigma_{\bar{x}} = \bar{x} - \mu_x \rightarrow \mu_x - z\sigma_{\bar{x}} = \bar{x}$$

i.e.

$$\bar{x} = \mu_x \pm z\sigma_{\bar{x}}$$

By symmetry of the standard normal distribution, we have



All we need to do now is substitute the appropriate values of μ , σ , and z .

By substitution the appropriate values of μ , σ , and z , we have

$$\begin{aligned} \bar{x} &= \mu_x \pm z\sigma_{\bar{x}} \\ &= 3.50 \pm 1.960 (0.270) \\ &= (2.9708, 4.0292) \end{aligned}$$

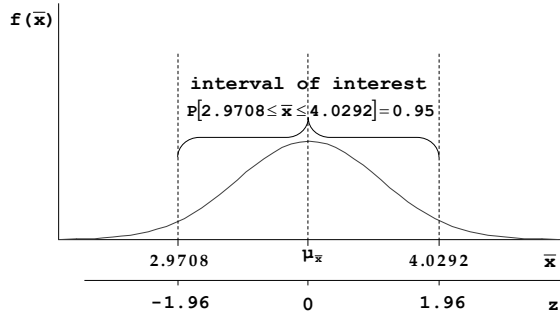
which can also be written

$$\begin{aligned} P(\mu_x - z_{\alpha/2}\sigma_{\bar{x}} \leq \bar{x} \leq \mu_x + z_{\alpha/2}\sigma_{\bar{x}}) &= 1 - \alpha, \text{ i.e.,} \\ P[3.50 - 1.960(0.270) \leq \bar{x} \leq 3.50 + 1.960(0.270)] &= 1 - 0.05 \\ \text{or} \\ P[2.9708 \leq \bar{x} \leq 4.0292] &= 0.95 \end{aligned}$$

So the (symmetric) interval of interest is (3.056, 3.944).

Note that the value of $1 - \alpha$ for a tolerance interval is frequently referred to as the 'tolerance level.'

By symmetry of the standard normal distribution, we have



We could do this for any level of tolerance and for asymmetric intervals.

4. The Sampling Distribution of the Sample Proportion - the probability distribution associated with all possible values of the sample proportion \bar{p} .

- the expected value of this distribution is p .
- the standard deviation (also called the standard error) of this distribution is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

if the population is 'infinite' in the statistical sense, i.e.

$$\frac{n}{N} \leq 0.05$$

Another Version of the Central Limit Theorem - when selecting a simple random sample from a population, the sampling distribution of \bar{p} can be approximated by a normal probability distribution as the sample size becomes large (here we will consider n to be sufficient to assure that the central limit theorem will force the potential values of \bar{p} to be normally distributed if

- $np \geq 5$ and
- $n(1-p) \geq 5$.

Example: If we toss a fair coin 35 times, what is the probability that we will get at least 20 heads?

The mean and standard error of \bar{p} are

$$E(\bar{p}) = p = 0.50$$

and

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.50(1-0.50)}{35}} = 0.084515$$

Since our sample can be considered 'sufficiently large,' i.e., $np \geq 5$ and $n(1-p) \geq 5$, we know that \bar{p} is approximately normally distributed.

Using the standard normal probability distribution, we have that

$$\begin{aligned} P\left(\bar{p} \geq \frac{20}{35}\right) &= P(\bar{p} \geq 0.5714) \\ &= P\left(\frac{\bar{p} - p}{\sigma_{\bar{p}}} \geq \frac{0.5714 - 0.50}{0.084515}\right) \\ &= P(z \geq 0.84) \\ &= P(-\infty \leq z \leq \infty) - P(-\infty \leq z \leq 0.84) \\ &= 1.000 - 0.7995 \\ &= 0.2005 \end{aligned}$$

Note that we could have used the binomial distribution to answer this question exactly (with a lot more effort) - the binomial distribution yields the (exact) answer of 0.24978.

The difference between the exact answer of 0.24978 and the approximate answer of 0.2005 from the normal distribution will get smaller as the sample size increases.

5. The Finite Population Correction Factor (fpcf) - if the sample size n is large relative to the size of the population N (n is at least 5% of N) the standard error must be adjusted downward. This is accomplished through multiplication of the standard error by the fpcf, which is

$$fpcf = \sqrt{\frac{N-n}{N-1}}$$

Thus, if the population is 'finite,' i.e., $\frac{n}{N} \geq 0.05$, the standard error of the sample mean becomes

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}} \right) \sqrt{\frac{N-n}{N-1}}$$

and standard error of the sample proportion becomes

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$
