# Sampling Distributions

A parameter is a numerical descriptive measure of the population (typically unknown)

A sample statistic is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

If we want to estimate a parameter of a population - say, the population mean $\mu$, there are a number of sample statistics that could be used for the estimate: e.g. sample mean $\bar{x}$ and sample medium $m$.

Q: which of these provides a better estimate of $\mu$?

Ex) die is tossed 3 times. Let $X$ = # of dots showing up after each roll.

population mean $\mu = \frac{(1+6)\times 3}{2\times 3} = \frac{1+6}{2} = 3.5$

suppose sample is $\{2, 2, 6\}$

Then $\bar{x} = \frac{2+2+6}{3} = 3.33$ and

sample median $m = 2$.

So for this sample of 3 measurements, the sample mean $\bar{x}$ provides an estimate that falls closer to $\mu$ than does the sample median.

Suppose another sample is $\{3, 4, 6\}$.

$\bar{x} = 4.33$ and $m = 4$.

This time $m$ is closer to $\mu$.

Point: Neither the sample mean nor the sample median will always fall closer to the population mean.

As random variables, sample statistics must be judged and compared on the basis of their probability distributions.

A <u>sampling distribution</u> of a sample statistic calculated from a sample of $n$ measurements is

the probability distribution of a statistic.

Ex) (external)

(A) from the table, you can see that $\bar{x}$ can assume the values 0, 1, 2, 3, 4, 5, 6, 8, 9, 12

$P(\bar{x}=0) = \frac{1}{27}$ (occurs in only {0,0,0} sample)

$\bar{x}=1$ occurs in 3 samples: (0,0,3), (0,3,0), (3,0,0)

$P(\bar{x}=1) = \frac{3}{27} = \frac{1}{9}$

The sampling distribution of $\bar{x}$ is its probability distribution:

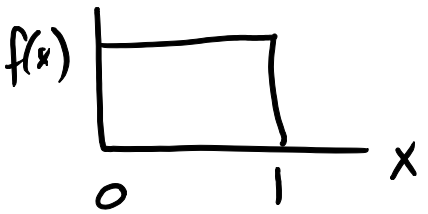
| $\bar{x}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| $P(\bar{x})$ | $\frac{1}{27}$ | $\frac{3}{27}$ | $\frac{3}{27}$ | $\frac{1}{27}$ | $\frac{3}{27}$ | $\frac{6}{27}$ | $\frac{3}{27}$ | $\frac{3}{27}$ | $\frac{3}{27}$ | $\frac{1}{27}$ |

(b) The median m can assume only 1 of 3 values: 0, 3, or 12. The value m=0 occurs in 7 different samples. m=3 in 13, m=12 in 7:

| m | 0 | 3 | 12 |
|---|---|---|---|
| P(m) | 7/27 | 13/27 | 7/27 |

In practice, we may choose to obtain an approximate sampling distribution for a statistic by simulating the sampling many times and recording the proportion of times different values of the statistic occur.

External R example Generate random values from uniform distribution on (0,1).




population mean = 0.5

we see that the values of the sample mean $\bar{x}$ cluster around $\mu$ to a greater extent than do the values of $m$.

Properties of Sampling Distributions: unbiasedness and minimum variance.

If a sampling distribution of a sample statistic has mean equal to the population parameter the statistic is intended to estimate, statistic is said to be

an unbiased estimate of the parameter.

Otherwise, it is a biased estimate of the parameter.

The standard deviation of a sampling distribution measures another important property of statistics, the spread of these estimates generated by repeated sampling.

Ex) We previously found the the sampling distribution of the sample mean $\bar{x}$ and the sample median $m$ for random samples of $n=3$ from the probability distribution:

| $x$ | 0 | 3 | 12 |
|---|---|---|---|
| $p(x)$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ |

sample mean

Show that $\bar{x}$ is an unbiased estimator for $m$ and that the sample median, $m$ is biased.

$$m = 0\left(\frac{1}{3}\right) + 3\left(\frac{1}{3}\right) + 12\left(\frac{1}{3}\right) = E[X] = \sum x p(x)$$

$$= 5$$

Recall the distributions of $\bar{x}$ and $m$:

| $\bar{x}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| $p(\bar{x})$ | 1/27 | 3/27 | 3/27 | 1/27 | 3/27 | 6/27 | 3/27 | 3/27 | 3/27 | 1/27 |

| $m$ | 0 | 3 | 12 |
|---|---|---|---|
| $p(m)$ | 7/27 | 13/27 | 7/27 |

$$E[\bar{x}] = \sum \bar{x} p(\bar{x})$$

$$E[\bar{x}] = 0\left(\frac{1}{27}\right) + 1\left(\frac{3}{27}\right) + 2\left(\frac{3}{27}\right) + \ldots + 12\left(\frac{1}{27}\right) = 5$$

Since $E[\bar{x}] = m$, we see that the sample mean $\bar{x}$ is an <u>unbiased</u> estimator of $m$.

$$E[m] = \sum m p(m) = 0\left(\frac{7}{27}\right) + 3\left(\frac{13}{27}\right) + 12\left(\frac{7}{27}\right)$$
$$= 4.56$$

Since $E[m] \neq m$, $m$ (the sample median) is a biased estimator of $m$.

what about the variance and std deviation?

$$\sigma_{\bar{x}}^2 = E\left\{(\bar{x} - E[\bar{x}])^2\right\} = \sum (\bar{x} - m)^2 p(\bar{x})$$

with $m = E[\bar{x}] = 5$.

$$\Rightarrow \sigma_{\bar{x}}^2 = (0-5)^2\left(\frac{1}{27}\right) + (1-5)^2\left(\frac{3}{27}\right) + (2-5)^2\left(\frac{3}{27}\right) +$$
$$+ \ldots + (12-5)^2\left(\frac{1}{27}\right) \simeq 8.667$$

$$\Rightarrow \sigma_{\bar{x}} \approx \sqrt{8.667} \approx 2.94$$

$$\sigma_m^2 = E\{([m - E[m])^2\} =$$

$$= \sum [m - E[m]]^2 p(m)$$

$$= (0 - 4.56)^2 \left(\frac{7}{27}\right) + (3 - 4.56)^2 \left(\frac{13}{27}\right) +$$

$$+ (12 - 4.56)^2 \left(\frac{7}{27}\right) = 20.9136$$

Let's look at more examples.

Overview: Two major activities of inferential statistics are (1): to use sample data to estimate values of population parameters, (2) to test hypothesis or claims made about population parameters.

Sampling distribution of a <u>statistic</u> (such as the sample mean), is the distribution of all values of that statistic (random variable) when all possible samples of the same size are taken from the same population.

There is often more than one choice of statistic to estimate population parameters.

Sample means $\bar{x}$ target the value of the population mean $\mu$. The distribution of the sample means tends to be a normal distribution.

The sampling distribution of the {mean, variance} is the distribution of sample {means, variances}, with all samples having the same sample size $n$ taken from the same population.

We may also talk about proportions and their sampling distribution.

The central limit theorem tells us that for a population with any distribution, the distribution of the sample means approaches a normal distribution as the sample size increases.

Let $\{x_1, x_2, \ldots, x_n\}$ constitute a random sample.

Then $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ and $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

sample mean  sample variance

Since these statistics are random variables, their values will vary sample to sample.

**Thm** | If $\{X_1, X_2, \ldots, X_n\}$ constitute a random sample from an infinite (or at least very large) population with mean $\mu$ and variance $\sigma^2$, then:

$$\boxed{E(\bar{x}) = \mu} \quad \text{and} \quad \boxed{\text{var}(\bar{x}) = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\Rightarrow E(\bar{x}) = \frac{1}{n}\left(E[x_1] + \ldots + E[x_n]\right) = \frac{1}{n}(\mu + \ldots + \mu)$$

$$= \frac{1}{n} n\mu = \mu$$

since $E[x_k] = \mu$ for all $k$ given that each $x_k$ has the same distribution as the population, having mean $\mu$.

**Corollary** | If the random variables $x_1, x_2, \ldots, x_n$ are independent and $y = \sum_{i=1}^{n} a_i x_i$, then:

$$\text{var}(y) = \sum_{i=1}^{n} a_i^2 \, \text{var}(x_i)$$

For the sample mean:

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

$$\Rightarrow \operatorname{Var}(\bar{X}) = \frac{1}{n^2}\operatorname{Var}(X_1) + \ldots + \frac{1}{n^2}\operatorname{Var}(X_n)$$

$$= n\left(\frac{1}{n^2}\sigma^2\right) = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Notice that the standard deviation of the distribution of $\bar{X}$ decreases as $n$ (the sample size) is increased. When $n$ becomes larger $\bar{X}_i$ become closer to $\mu$, the quantity they are expected to estimate.
$\bar{X}_i$ sample mean from $i$-th sample of size $n$.

Ex) Suppose an elevator has a maximum capacity of 16 passangers with a total weight of 2500 lb.

Assume that at some point 16 males are in the elevator. We like to find the probability that the elevator is overloaded.

Assume male weights follow a normal distribution

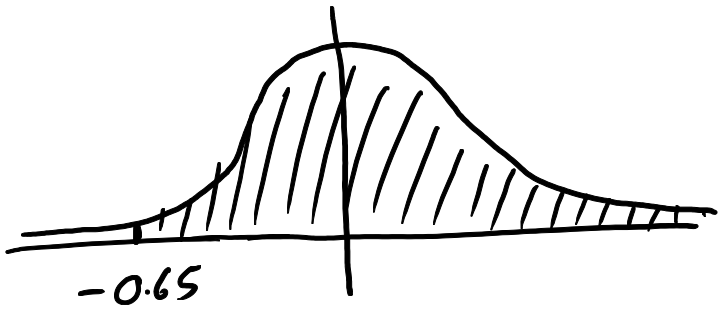
with a mean of 182.9 lb and a standard deviation of 40.8 lb.

(A) Find the probability that one randomly selected male weighs over 156.25 lb.

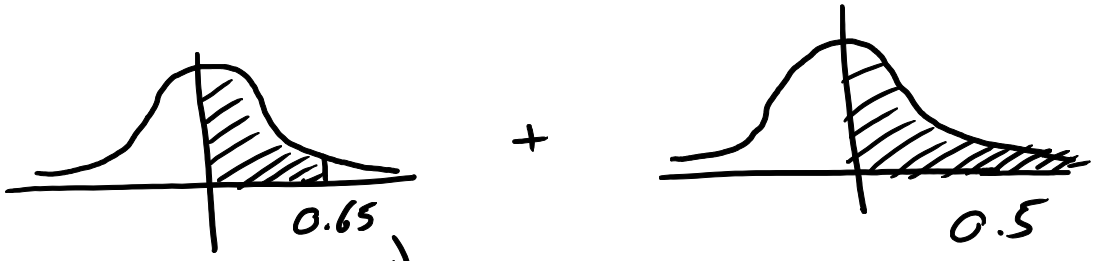
$$z = \frac{x - \mu}{\sigma} \text{ is standard normal}$$

$$P(x > 156.25) = P\left(\frac{x-\mu}{\sigma} > \frac{156.25-\mu}{\sigma}\right)$$

$$= P\left(z > \frac{156.25 - 182.9}{40.8}\right) = P(z > -0.65)$$




This area is the same as these two areas:




A(0.65)

↑ use table

$$0.2422 + 0.5 = 0.7422$$

(B) Find the probability that a sample of 16 males have a mean weight greater than 156.25 lb.

The distribution of sample means is assumed to be normal.

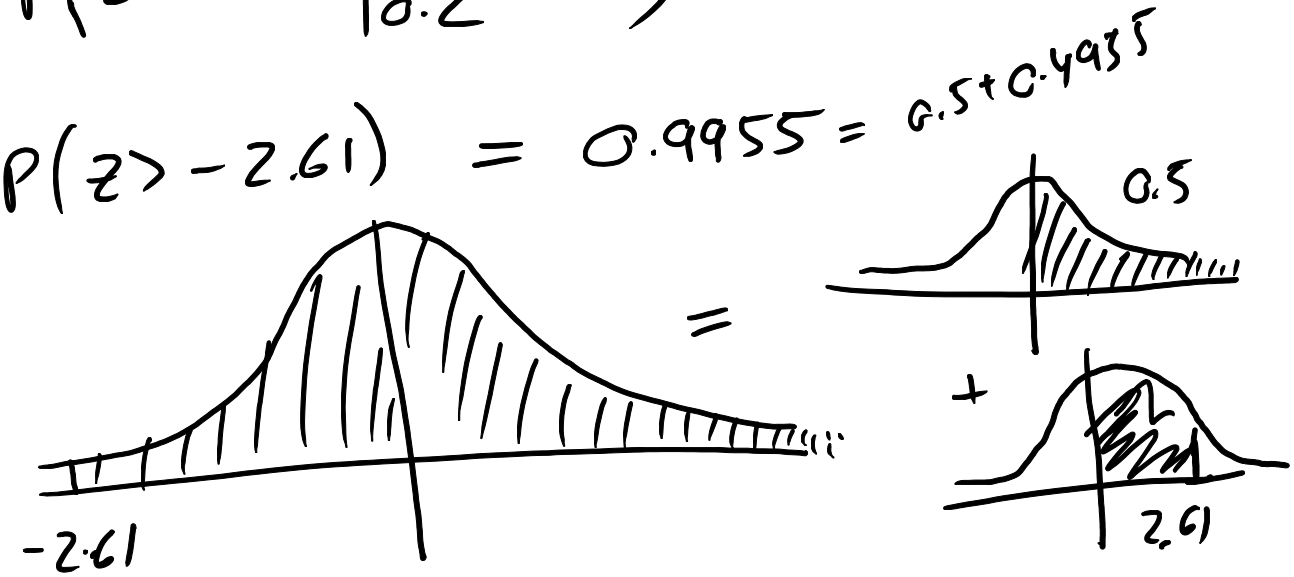
$$\mu_{\bar{x}} = \mu = 182.9$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{40.8}{\sqrt{16}} = 10.2$$

$$P(\bar{x} > 156.2) = P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{156.2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P\left(z > \frac{156.2 - 182.9}{10.2}\right)$$

$$= P(z > -2.61) = 0.9955 = 0.5 + 0.4955$$




P( Any given male weighs > 156.25 lb) = 0.7432

P( mean of sample of 16 males > 156.25 lb) = 0.9955

For 16 males, high chance that 16×156.25 will be exceeded (safe capacity exceeded).

$\mu = 80, \sigma = 5$ population quantities

Ex) Suppose we have selected a random sample from a population with mean equal to 80 and std deviation equal to 5. Assume population is normally distributed.

Find the probability that the sample mean $\bar{x}$ is greater than 82. $n = 25$ observations.

$E(\bar{x}) =$

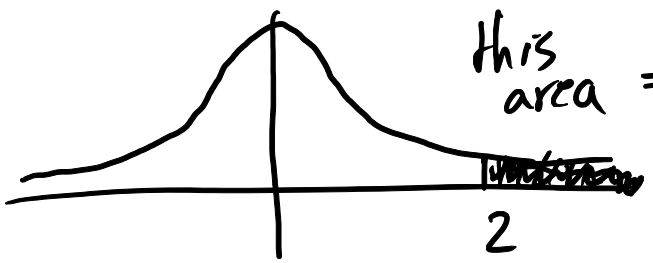
$= \mu_{\bar{x}} = \mu = 80$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$

$$P(\bar{x} > 82) = P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{82 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

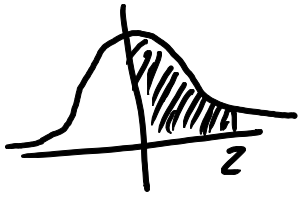
($\bar{x}$ is normally distributed)

$$= P\left(z > \frac{82 - 80}{1}\right) = P(z > 2)$$

$$= 0.5 - .4772 = .0228$$




this area $= 0.5 -$



Central limit theorem for sample means:

Let $\{X_1, X_2, \ldots, X_n\}$ be a simple random sample from a population with mean $\mu$ and variance $\sigma^2$.

Let $\bar{x} = \frac{\sum x_i}{n}$ be the sample mean

Then if $n$ is sufficiently large, the sample mean has an approximately normal distribution with

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

($\bar{x}$ is an unbiased estimator of $\mu$)

$\sigma_{\bar{x}}$ is often called the std. error of the mean → measures variability of the sample mean.

Note: How quickly the distribution of the sample mean approaches normality depends on distribution of population. $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p\}$

Rule of thumb: $n \geq 30$.

(Inferring something about a population from a sample)

Ex) Based on CLT, what is the probability that the error (in estimation of the population mean $\mu$) will be less than 5, when the mean of a random sample of size $n=64$ is used to estimate the mean of an infinite population with $\sigma = 20$?

$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

$z$ has standard normal distribution (approx.)

notice that we must divide by $\sigma_{\bar{x}} \neq \sigma$.

We want $P(|\bar{x}-\mu| < 5)$

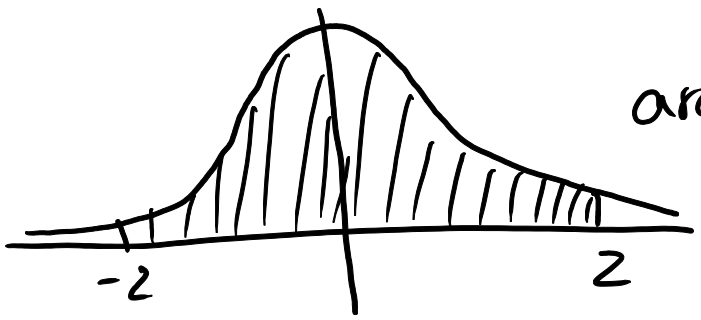
$$P(|\bar{x}-\mu|<5) = P[-5 < (\bar{x}-\mu) < 5]$$

$$= P\left[\frac{-5}{\sigma_{\bar{x}}} < \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} < \frac{5}{\sigma_{\bar{x}}}\right]$$

$$= P\left[\frac{-5}{20/\sqrt{64}} < z < \frac{5}{20/\sqrt{64}}\right]$$

Now $\frac{5}{20/\sqrt{64}} = \frac{5}{20/8} = \frac{5}{5/2} = 2$

$\Rightarrow P(|\bar{X} - \mu| < 5) = P(-2 < z < 2) = 0.9544$




area $= 2A(2)$
$= 2 \times 0.4772$
$= 0.9544$

Ex) (inferring something about a sample given population statistics)

The number of flaws on DVDs produced by a company has the following probability distribution:

| X | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| P(x) | 0.60 | 0.25 | 0.10 | 0.05 |

Suppose 36 DVDs are sampled from this population. What is the probability that the mean number of flaws per DVD in this sample is less than 0.5? ($\Leftrightarrow$ Prob. that # of flaws in 36 DVDs is less than 18?)

$= E[x] = \Sigma x p(x)$

$\mu = 0(0.60) + 1(0.25) + 2(0.10) + 3(0.05) = 0.6$

$\mu$ is the population mean

$\sigma^2 = E[(x - E[x])^2] =$

$= \Sigma (x-\mu)^2 p(x)$

$= (0-0.6)^2(0.60) + (1-0.6)^2(0.25)$

$+ (2-0.6)^2(0.10) + (3-0.6)^2(0.05)$

$= (0.6)^3 + (0.4)^2(0.25) + (1.4)^2(0.1) + (2.4)^2(0.05)$

$= 0.74$

$\sigma = \sqrt{0.74}$ is the population std. deviation

Since $n$ is large enough ($n > 30$), we can apply the CLT. Sample mean $\bar{x}$ is approx. normally distributed with

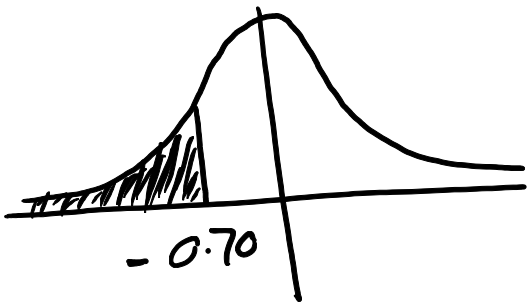
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.74}}{\sqrt{36}} \approx 0.143$$

and $\mu_{\bar{x}} = \mu = 0.6$

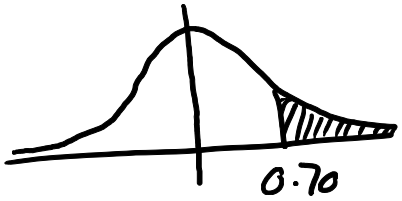
we want $P(\bar{x} < 0.5)$

$$P(\bar{x} < 0.5) = P\left(\frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} < \frac{0.5 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P\left(z < \frac{0.5 - 0.6}{0.143}\right) = P(z < -0.70)$$




this area is the same as:




$$= 0.5 - A(0.70) = 0.5 - 0.2580 = \boxed{0.2420}$$

where $A(0.70) = P(0 < z < 0.70)$

## Approximation of a binomial distribution with a normal distribution

If the conditions $np \geq 5$ and $nq \geq 5$ are satisfied, then the probabilities from a binomial distribution can be approximated well by using a normal distribution with:

$$\mu = np \text{ and } \sigma = \sqrt{npq}$$

**continuity correction** When we use a normal distribution as an approx. to the binomial distribution (which is discrete), a continuity correction is made to a discrete whole number $x$ by using the interval from $x-0.5$ to $x+0.5$.

Ex) In 431 coin tosses, find the probability of getting at least 235 heads.

$X$ counts # of heads

$X$ a binomial random variable

$p=0.5$, $q=0.5$, $n=431$ (# of trials)

$$P(X \geq 235) = 1 - P(X \leq 234)$$

$$= 1 - [P(x=0) + P(x=1) + \ldots + P(x=234)]$$

$$= 1 - \sum_{k=0}^{234} \binom{431}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{431-k}$$

$$\mu = np = 431\left(\frac{1}{2}\right) = 215.5 \geq 5$$

$$nq = np \geq 5$$

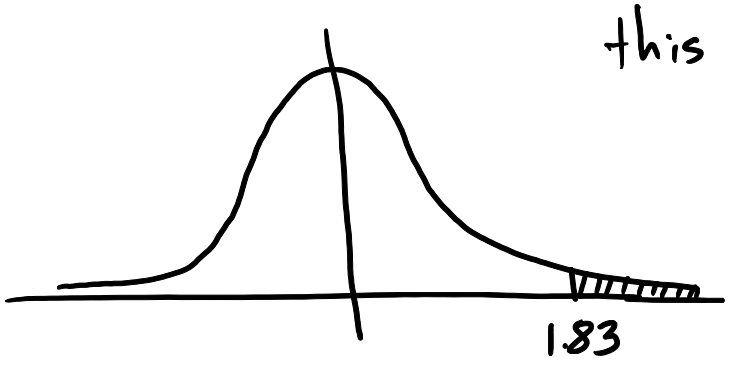
$$\sigma = \sqrt{npq} = \sqrt{431\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} \approx 10.38$$

$$P(X_{binom} \geq 235) \approx P(X_{normal} \geq 234.5)$$

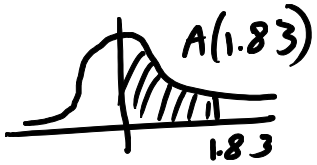
(subtract 0.5 from 235 for continuity correction)

$$= P\left(\frac{X-\mu}{\sigma} \geq \frac{234.5-215.5}{10.38}\right)$$

$$= P(z \geq 1.83)$$




this area $= 0.5 - A(1.83)$




$$= 0.5 - 0.4664 = 0.0336$$

## Estimating a population proportion

The sample proportion as an estimate of the population proportion.

A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter.

A confidence interval is the probability $(1-\alpha)$ that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times.