Sampling theory

Random sampling, sampling distributions, standard error, Central Limit Theorem, normal distribution, probability, confidence intervals

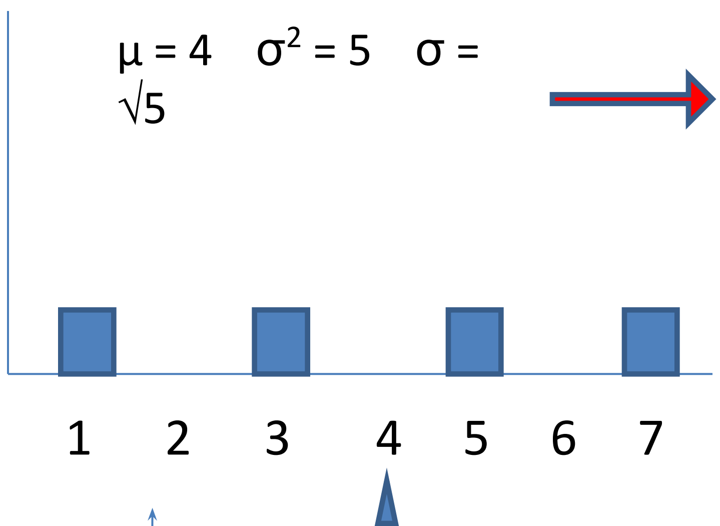
Sampling random variables

- Randomly draw a sample from a population distribution
  - Each case has an equal chance of being selected
  - Each selection is independent of the others
    - therefore, cases are drawn with replacement
- Illustrate using a discrete variable with only a few values {1, 3, 5, 7} in the population
  - same principles apply to continuous variables

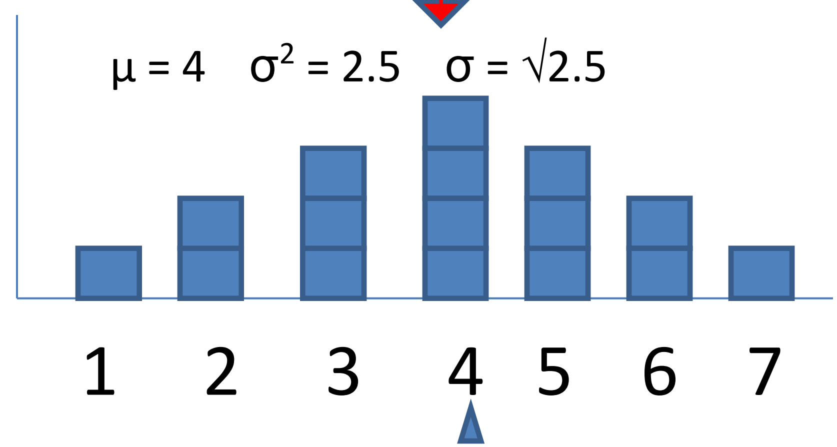
Samples of size n = 2

- Draw from the population every possible sample consisting of 2 scores; sum the 2 scores & divide by 2; plot the means

(1+1)/ 2=1	(1+3)/2=2	(1+5)/2=3	(1+7)/2=4
(3+1)/2=2	(3+3)/2=3	(3+5)/2=4	(3+7)/2=5
(5+1)/2=3	(5+3)/2=4	(5+5)/2=5	(5+7)/2=6
(7+1)/2=4	(7+3)/2=5	(7+5)/2=6	(7+7)/2=7



Population



Sampling Distribution

Sampling distribution

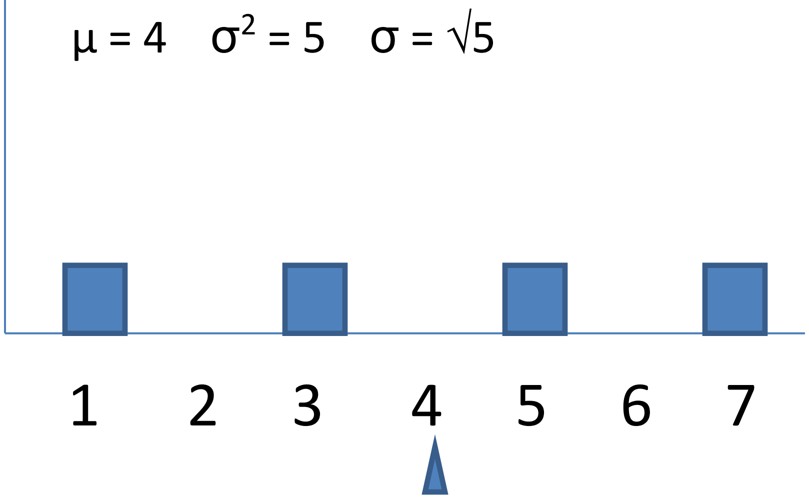
- Distribution of a sample statistic for samples of size *n* drawn from a population distribution
- Sampling distribution of the mean
  - Sum each sample's scores, divide by the *n*

Mean = μ

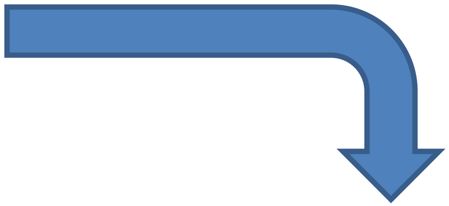
Variance = σ^2 / n S.D. = $\sqrt{\frac{\sigma^2}{n}}$ $\frac{\sigma}{\sqrt{n}}$

Probability of a sampling outcome

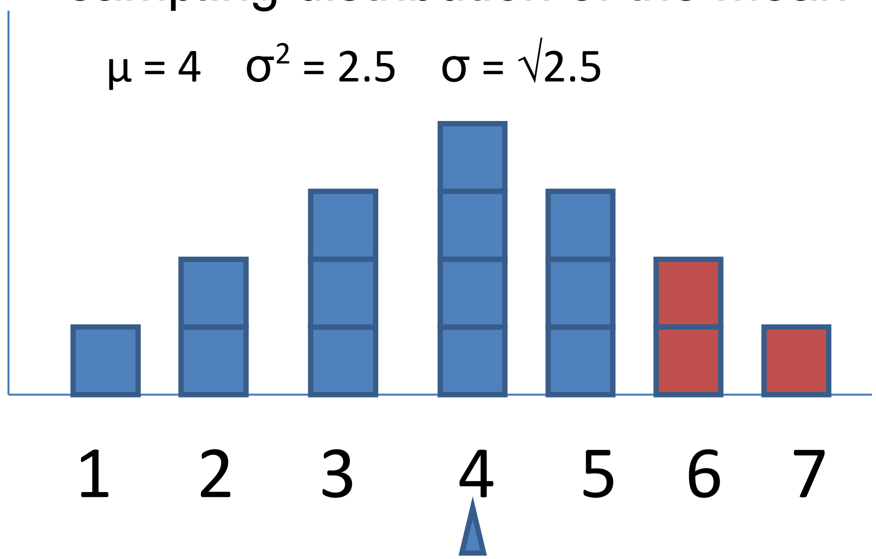
population



draw samples, sum each & divide by *n*



sampling distribution of the mean



$$P(A) = \frac{\#\, outcomes\ classified\ as\ A}{total\ \#\ of\ \ possible\ outcomes}$$

$$P(\overline{X} \geq \mu + 2) = \frac{3}{16} = 0.1875$$

Sampling distribution of the mean

- Mean of the distribution = μ
  - the mean of sample means is μ
- Variance of the distribution $$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$
  - less variable than source population
- Standard deviation of the distribution $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
  - Also known as "standard error of the mean" (SE)
- Shape differs from source population
  - more values near mean
  - bell shaped

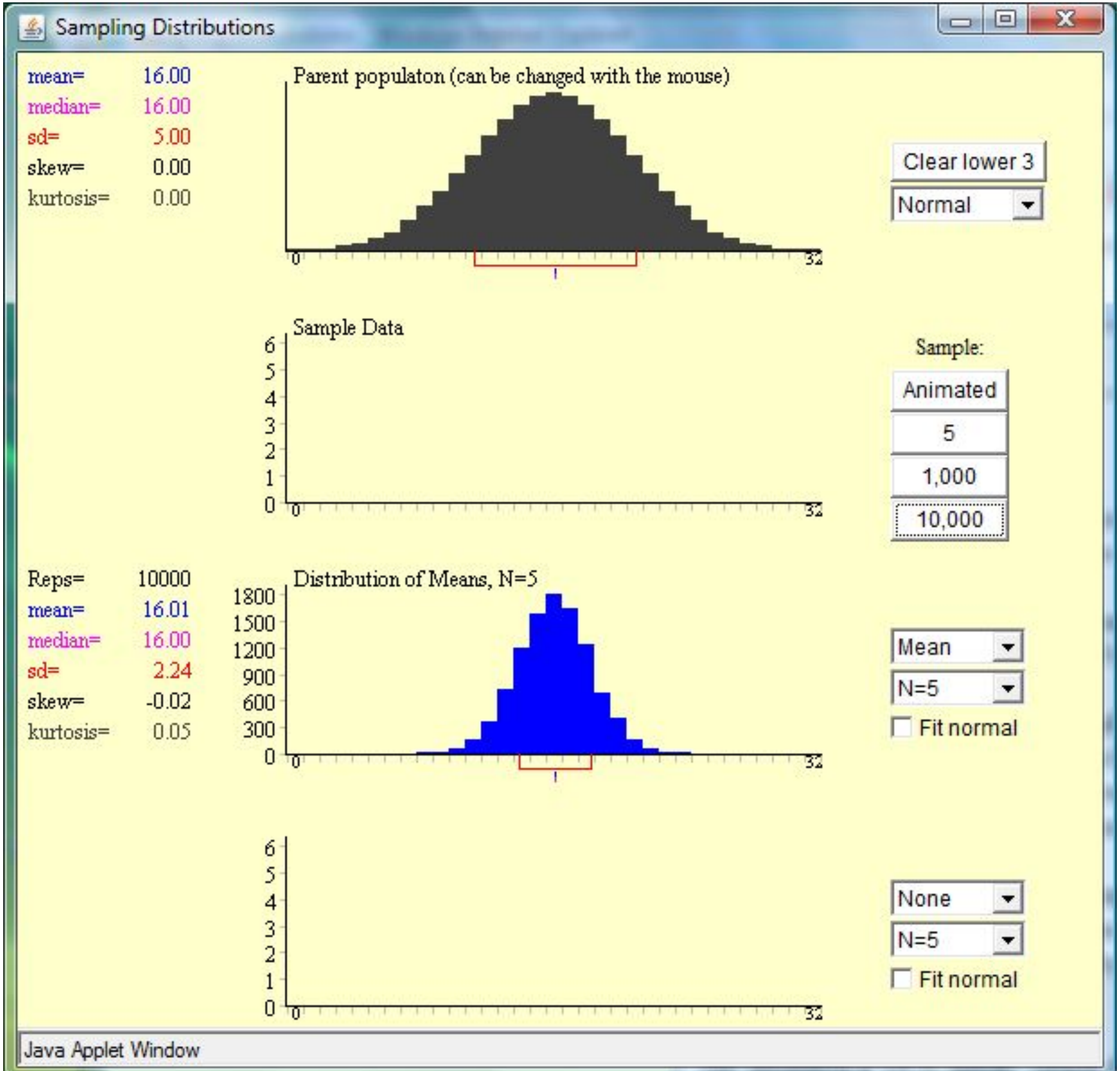
Central Limit Theorem (CLT)

- As the sample size *n* increases, the shape of the sampling distribution of the mean approaches the shape of the normal distribution (a.k.a. “the bell curve”, “the Gaussian distribution”)
- Importance: the normal distribution allows us to map distance from the mean to probability of occurrence, *even if the population distribution is not normal, provided that the sample size, n, is not small (> approx. 30)*

Online simulation

- Uses random number generator to sample from population
- Empirical (as opposed to theoretical) sampling distributions

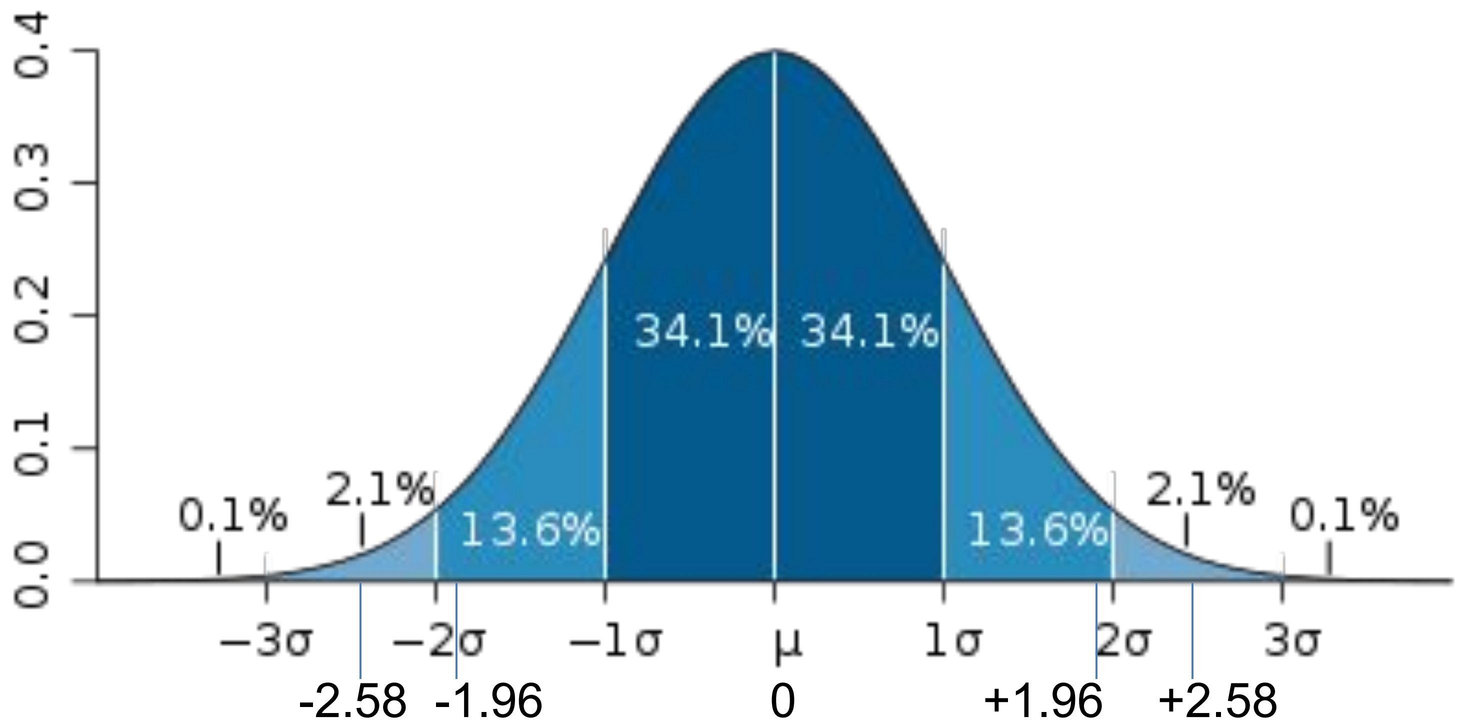
http://onlinestatbook.com/stat_sim/sampling_dist/index.html


Sampling Distributions
mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00
Parent populaton (can be changed with the mouse)
Clear lower 3
Normal
Sample Data
Sample:
Animated
5
1,000
10,000
Reps= 10000
mean= 16.01
median= 16.00
sd= 2.24
skew= -0.02
kurtosis= 0.05
Distribution of Means, N=5
Mean
N=5
Fit normal
None
N=5
Fit normal
Java Applet Window

A do-it-yourself sampling distribution

```
> rnorm(16, 0, 1)
 [1] -1.13505779  0.74416008  0.03917791  0.41535716 -1.31979649
-0.98551010  1.35561128
 [8]  2.87106735  1.76864786 -0.94445105 -1.00517080 -0.07183120
1.20544913  0.67444393
[15] -0.66605983 -0.13354738
> mean(rnorm(16, 0, 1))
[1] 0.101037
> x <- replicate(10000, mean(rnorm(16, 0, 1)))
> hist(x, xlim = c(-1, 1), freq = F)
> summary(x)
      Min.     1st Qu.         Median        Mean      3rd Qu.     Max.
-1.02900   -0.17160   -0.00317   -0.00083   0.16550    0.97720
> sd(x)
[1] 0.2496625
> library(moments)
> skewness(x)
[1] 0.05395557
> kurtosis(x) - 3
[1] -0.02523182
```

The normal distribution probability density function



$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The normal distribution

- Extremely common in nature
  - height of adult males; height of corn plants in a field
  - IQ scores (by design)
  - Sum of many random variables
- But not all populations are normal; e.g., reaction times (cannot be < 0)
- Sampling distribution of mean will be (approximately) normal if population is normal ***OR*** sample size is large (> 30)

Area underneath the normal curve

z	z to mean	smaller area	larger area
0.00	0.0000	0.5000	0.5000
0.10	0.0398	0.4602	0.5398
0.20	0.0793	0.4207	0.5793
0.30	0.1179	0.3821	0.6179
0.40	0.1554	0.3446	0.6554
0.50	0.1915	0.3085	0.6915
0.60	0.2257	0.2743	0.7257
0.70	0.2580	0.2420	0.7580
0.80	0.2881	0.2119	0.7881
0.90	0.3159	0.1841	0.8159
1.00	0.3413	0.1587	0.8413
1.10	0.3643	0.1357	0.8643
1.20	0.3849	0.1151	0.8849
1.30	0.4032	0.0968	0.9032
1.40	0.4192	0.0808	0.9192
1.50	0.4332	0.0668	0.9332
1.60	0.4452	0.0548	0.9452
1.70	0.4554	0.0446	0.9554
1.80	0.4641	0.0359	0.9641
1.90	0.4713	0.0287	0.9713
2.00	0.4772	0.0228	0.9772

z	z to mean	smaller area	larger area
2.00	0.4772	0.0228	0.9772
2.10	0.4821	0.0179	0.9821
2.20	0.4861	0.0139	0.9861
2.30	0.4893	0.0107	0.9893
2.40	0.4918	0.0082	0.9918
2.50	0.4938	0.0062	0.9938
2.60	0.4953	0.0047	0.9953
2.70	0.4965	0.0035	0.9965
2.80	0.4974	0.0026	0.9974
2.90	0.4981	0.0019	0.9981
3.00	0.4987	0.0013	0.9987
3.10	0.4990	0.0010	0.9990
3.20	0.4993	0.0007	0.9993
3.30	0.4995	0.0005	0.9995
3.40	0.4997	0.0003	0.9997
3.50	0.4998	0.0002	0.9998
3.60	0.49984	0.00016	0.99984
3.70	0.49989	0.00011	0.99989
3.80	0.49993	0.00007	0.99993
3.90	0.49995	0.00005	0.99995
4.00	0.49997	0.00003	0.99997

Sampling distribution of the mean

- Fortunately, we do not need to construct the sampling distribution of the mean in order to use its properties.
- If population parameters μ and σ are known:
  - if the population is normal, we can use CLT to answer questions about probability of obtaining certain sample means, or
  - if the population is non-normal, we can use CLT so long as sample size is large.

The `pnorm` function

`pnorm`: area under the normal curve

```
pnorm(q, mean = 0, sd = 1,
      lower.tail = TRUE)
```

where `q` is a score or a vector of scores and `lower.tail = TRUE` selects tail of distribution.

- `pnorm(67, 100, 20)` gives the probability of score < 67 when $\mu = 100$ and $\sigma = 20$.
- `pnorm(2.1, lower.tail = F)` gives the probability of score > 2.1 when $\mu = 0$ and $\sigma = 1$.

Probability of an individual with a particular score

IQ scores are distributed approximately normally with $\mu = 100$ and $\sigma = 15$. What is the probability of randomly selecting an individual with an IQ > 118?

```
> pnorm(118, 100, 15,
        lower.tail = F)
[1] 0.1150697
```

Probability of an individual with a score within a particular range

IQ scores are distributed approximately normally with $\mu = 100$ and $\sigma = 15$. What is the probability of randomly selecting an individual with an IQ > 79 & < 94 ?

```
> areas <- pnorm(c(79, 94),
                  100, 15)
> areas[2] - areas[1]
[1] 0.2638216
```

Probability of a sample with a mean of a particular value

IQ scores are distributed approximately normally with $\mu = 100$ and $\sigma = 15$. What is the probability of randomly selecting a sample of 4 individuals with a mean IQ greater than 118?

Recall that: $SE = \sigma/\sqrt{n} = 15 / 2 = 7.5$

```
> pnorm(118, 100, 15 / sqrt(4),
        lower.tail = F)
[1] 0.008197536
```

The `qnorm` function

`qnorm`: returns score(s) delimiting area(s) under the normal curve

```
qnorm(p, mean = 0, sd = 1,
        lower.tail = TRUE)
```

where `p` is an area (a probability) or a vector of areas (probabilities) and `lower.tail = TRUE` selects tail of distribution.

- `qnorm(.25, 100, 20)` gives the 25th percentile when $\mu = 100$ and $\sigma = 20$.
- `qnorm(.1, lower.tail = F)` gives the 90th percentile when $\mu = 0$ and $\sigma = 1$.

Finding a range of scores within which a particular % of sample means fall

Between what 2 values of IQ would we expect the means of random samples of size n=36 to fall 95% of the time?

Recall that: $SE = \sigma / \sqrt{n} = 15 / 6 = 2.5$

```
> qnorm(c(.025, .975), 100,
        15 / sqrt(36))
[1]  95.10009  104.89991
```

A final example

- In the lexical decision task, subjects must decide if a string of letters is a word, and their time to decide is measured as the dependent variable
  - Q: If lexical decision reaction times have $\mu = 759$ ms and $\sigma = 176$ ms, then between what 2 values would we expect the means of random samples of n=10 to fall 95% of the time?
  - A: Distributions of reaction times are usually skewed. The data should be examined visually and compared to the normal distribution. With a small sample size ($n = 10$), use of the normal distribution is probably not justified because of the skew.

We have assumed knowledge of μ & σ, but population parameters are rarely known

- Many population distributions have never been studied before, such as:
  - a new survey scale in social or clinical psychology
  - acceptability judgments for syntactic constructions, or
  - effects of a new method of teaching statistics.
- Even slight changes in presentation or measurement conditions can alter mean and variance of data.

Suppose σ is known but μ is unknown

- Q: If a sample of size n=16 with a mean of Xbar = 20 is drawn from a normal population with σ = 10, what is the value of μ?
- A: Construct a confidence interval (CI), a region within which, we believe, μ is located.

<u>for a 95% CI</u>, find lower limit that is 1.96 SE below Xbar

find upper limit that is 1.96 SE above Xbar

<u>for a 99% CI</u>, find lower limit that is 2.58 SE below Xbar

find upper limit that is 2.58 SE above Xbar

Confidence Intervals (CIs)

- A: 95% CI (continued)

  sample Xbar = 20, n = 16,

  population σ = 10

SE = 10 / 4 = 2.5

```
> qnorm(c(.025, .975), 20,
        10 / sqrt(16))
[1] 15.10009  24.89991
```

- A: 99% CI (continued)

  sample Xbar = 20, n = 16,

  population σ = 10

SE = 10 / 4 = 2.5

```
> qnorm(c(.005, .995), 20,
        10 / sqrt(16))
[1] 13.56043  26.43957
```

Interpretation of a confidence interval

- 95% of the time that the procedure for constructing a 95% CI is followed, μ is within the CI,
- 99% of the time that the procedure for constructing a 99% CI is followed, μ is within the CI,
- and so on.

Online simulations:

http://onlinestatbook.com/stat_sim/conf_interval/index.html

http://wise1.cgu.edu/vis/ci_creation/

Confidence intervals: demonstration