



LECTURE 3 – INFERENCE STATISTICS

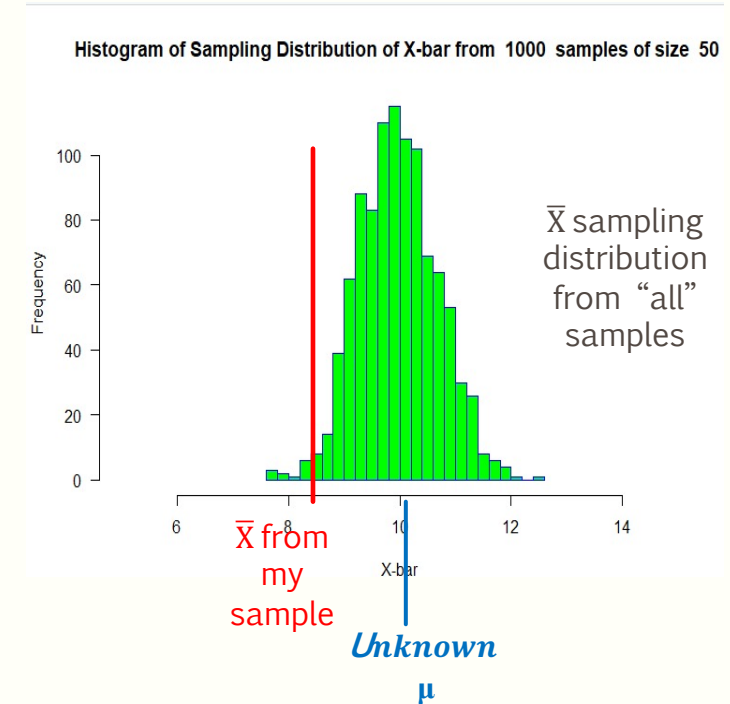
Part B – Sampling Distribution of \bar{X}

The Statistical Inference Problem

- Thus far, we assumed that we knew the population distribution and we simulated many samples from that distribution and obtained the sampling distribution of a statistic.
- However, in practice we:
 - We don't know the population distribution
 - We usually take only *one sample*
 - *Our interest is knowing a single population parameter* (such as the unknown population mean, μ)
- In other words, we would like to infer the value of a single (unknown) population parameter, from a single statistic, calculated from a single sample.
- This is the *inference problem* in statistics.
- We would like to, for example, use a single \bar{X} to get an estimate of μ as well as understand how good our estimate is.

The Statistical Inference Problem

- We then have the following question:
 - If we can have only one sample, and therefore one \bar{X} , how do we know how “close” it is to the population parameter μ ?
- A different way to look at the same question is:
 - How does my \bar{X} from a single sample compare with the “ \bar{X} from all the possible samples (of the same size as my sample)”?
- Remember that “ \bar{X} from all the possible samples (of the same size as my sample)” is the same as the sampling distribution of \bar{X} !
- So, clearly, to answer these questions, **you need to know the sampling distribution of \bar{X}** .
- We cannot take all possible samples and we may or may not know the population distribution.

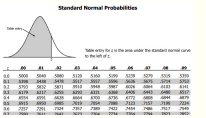


Sampling Distribution of \bar{X} - Four Cases

- Case 1: Population is (known/assumed) Normal (unknown μ , known σ)
 - Sampling distribution of \bar{X} from a sample of size n is: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$; Standard Error is $\frac{\sigma}{\sqrt{n}}$
- Case 2: Population is (known/assumed) Normal (unknown μ , unknown σ)
 - $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$ if sample size ≥ 30 , where s is the sample standard deviation; Standard Error is $\frac{s}{\sqrt{n}}$
 - $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})(n-1)$ degrees of freedom if sample size < 30 . Use **t**-distribution; Standard Error is $\frac{s}{\sqrt{n}}$
- Case 3: Population is unknown (unknown μ , known σ)
 - $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Standard Error is $\frac{\sigma}{\sqrt{n}}$; This is based on the **Central Limit Theorem**.
 - If sample size $\ll 30$, we really have to assume the population distribution. To avoid this, collect a larger sample.
- Case 4: Population is unknown (unknown μ , unknown σ)
 - Bootstrapping

Using Case 1: Population is (known/assumed) Normal (unknown μ , known σ)

Practice with Table



Standard Normal Probabilities

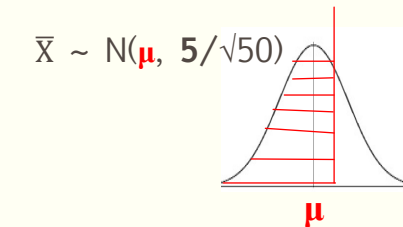
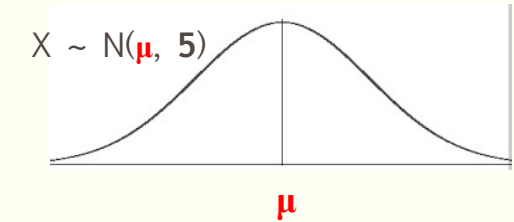
Table giving the area under the standard normal curve to the left of Z.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5518	.5558	.5597	.5637	.5677	.5717	.5757
0.2	.5797	.5837	.5877	.5917	.5957	.5997	.6037	.6077	.6117	.6157
0.3	.6197	.6237	.6277	.6317	.6357	.6397	.6437	.6477	.6517	.6557
0.4	.6597	.6637	.6677	.6717	.6757	.6797	.6837	.6877	.6917	.6957
0.5	.6997	.7037	.7077	.7117	.7157	.7197	.7237	.7277	.7317	.7357
0.6	.7397	.7437	.7477	.7517	.7557	.7597	.7637	.7677	.7717	.7757
0.7	.7797	.7837	.7877	.7917	.7957	.7997	.8037	.8077	.8117	.8157
0.8	.8197	.8237	.8277	.8317	.8357	.8397	.8437	.8477	.8517	.8557
0.9	.8597	.8637	.8677	.8717	.8757	.8797	.8837	.8877	.8917	.8957
1.0	.8997	.9037	.9077	.9117	.9157	.9197	.9237	.9277	.9317	.9357
1.1	.9397	.9437	.9477	.9517	.9557	.9597	.9637	.9677	.9717	.9757
1.2	.9797	.9837	.9877	.9917	.9957	.9997				

- $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$; Standard Error is $\frac{\sigma}{\sqrt{n}}$

- Example:

- Suppose I take a sample of size 50 from a population that is assumed normal $N(\mu, 5)$ and my estimator is the sample mean \bar{X} with a value of 10.3, what is my best (point) estimate of the unknown population mean μ ?
 - The answer, of course, is 10.3. this is the point estimate of the unknown population mean.
- Do I have any more information that I can get? The answer is: Yes.
- If my sample mean $\bar{X} = 10.3$, is it more likely to have come from a population with $\mu = 10$ or a population with $\mu = 11$?



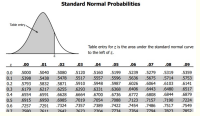
```
> pnorm(10.3, 10, 5/sqrt(50))  
[1] 0.6643134
```

```
> pnorm(10.3, 11, 5/sqrt(50))  
[1] 0.1610994
```

- The sampling distribution of $\bar{X} \sim N(\mu, 5/\sqrt{n})$ or $\sim N(\mu, 0.7071)$, since $n = 50$.
- Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 10) = P(Z \leq \frac{10.3 - 10}{0.7071}) = 0.664$
- Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 11) = P(Z \leq \frac{10.3 - 11}{0.7071}) = 0.161$
- Thus, my sample (mean) is more likely to have come from a population with $\mu = 10$ or a population with $\mu = 11$

Case 2: Population is (known/assumed) Normal (unknown μ , unknown σ) and sample size $n \geq 30$.

Practice with Table



- We don't know the population standard deviation, so we are going to use the sample standard deviation s .
- If the sample size $n \geq 30$, then the sampling distribution $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$, where $\frac{s}{\sqrt{n}}$ is the standard error.
- Assume that we took a sample and had the following sample statistics:
 - $\bar{X} = 10.3$, $s = 5.5$ and $n = 50$.
 - The sampling distribution of $\bar{X} \sim \text{Normal}(\mu, \frac{5.5}{\sqrt{50}}) = \text{Normal}(\mu, 0.7778)$
 - $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 10) = P(Z \leq \frac{10.3 - 10}{0.7778}) = 0.650$. There is a 0.650 probability that $\bar{X} \leq 10.3$ if our assumption $\mu = 10$ is correct.

```
> pnorm(10.3, 10, 0.7778)
[1] 0.6501418
```
 - Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 11) = P(Z \leq \frac{10.3 - 11}{0.7778}) = 0.184$. There is a 0.184 probability that $\bar{X} \leq 10.3$ if our assumption $\mu = 11$ is correct.

```
> pnorm(10.3, 11, 0.7778)
[1] 0.184067
```
- Thus, my sample (mean) is more likely to have come from a population with $\mu = 10$ or a population with $\mu = 11$

Case 2: Population is (known/assumed) Normal (unknown μ , unknown σ) and sample size $n < 30$.

- If the sample size $n < 30$, then the sampling distribution the standardized value of \bar{X} , namely $(\bar{X} - \mu)/(\frac{s}{\sqrt{n}})$ follows a *Student's t-distribution*. It was developed by William Sealy Gosset under the pseudonym *Student*.
- The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.
- The t-distribution is characterized by $(n-1)$ degrees of freedom, where n is the sample size.
- At around a sample size of 30, the t-table coincides with the standard normal table.
- Assume that we took a sample and had the following sample statistics:
 - $\bar{X} = 10.3$, $s = 5.3$ and $n = 25$. So, standard error $\frac{s}{\sqrt{n}} = \frac{5.3}{\sqrt{25}} = 1.06$
 - Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 10) = P(t \leq (10.3 - 10)/(1.06) = P(t \leq 0.2830)$ with 24 degrees of freedom.

```
> pt(0.2830, 24, lower.tail = TRUE, log.p = FALSE)
[1] 0.6101983
```

= 0.6102. There is a 0.6102 probability that $\bar{X} \leq 10.3$ if our assumption $\mu = 10$ is correct.
 - Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 11) = P(t \leq (10.3 - 11)/(1.06) = P(t \leq -0.6604)$ with 24 degrees of freedom.

```
> pt(-0.6604, 24, lower.tail = TRUE, log.p = FALSE)
[1] 0.2576431
```

= .2576. There is a 0.2576 probability that $\bar{X} \leq 10.3$ if our assumption $\mu = 11$ is correct.
 - Thus, my sample (mean) is more likely to have come from a population with $\mu = 10$ or a population with $\mu = 11$

The t-table

t Distribution						
α						
Degrees of freedom	.005 (one tail) .01 (two tails)	.01 (one tail) .02 (two tails)	.025 (one tail) .05 (two tails)	.05 (one tail) .10 (two tails)	.10 (one tail) .20 (two tails)	.25 (one tail) .50 (two tails)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.132	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.540	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.787	2.485	2.060	1.708	1.316	.684
26	2.779	2.479	2.056	1.706	1.315	.684
27	2.771	2.473	2.052	1.703	1.314	.684
28	2.763	2.467	2.048	1.701	1.313	.683
29	2.756	2.462	2.045	1.699	1.311	.683
Large (∞)	2.575	2.327	1.960	1.645	1.282	.675

We cannot use the t-table for the problems we just did, but we will use it later for other problems.

Case 3: Population is **unknown** (unknown μ , known σ)

- Even if we don't know the population distribution, under two special conditions, namely:
 1. We know the population standard deviation σ
 2. The sample size is large enough (generally 30 or more)
- The **Central Limit Theorem or CLT (for \bar{X})** says:
 - Sampling distribution of \bar{X} is: $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$
- This is a really powerful result, because it applies even to unknown populations.
- It also says:
 - $E(\bar{X}) = \mu$ (i.e., \bar{X} is an *unbiased* estimator of the unknown population parameter μ we are trying to estimate)
 - $SD(\bar{X}) = \text{Standard Error} = \frac{\sigma}{\sqrt{n}}$ is a measure of *precision* of our estimator \bar{X} in estimating μ .
- **Note:**
 - The CLT shown here *only applies* when we are using \bar{X} estimate μ .
 - When we are estimating the sampling distribution of *other* estimators, to estimate μ or some other population parameter, we cannot use CLT.

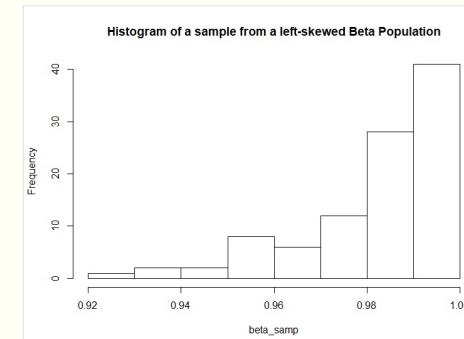
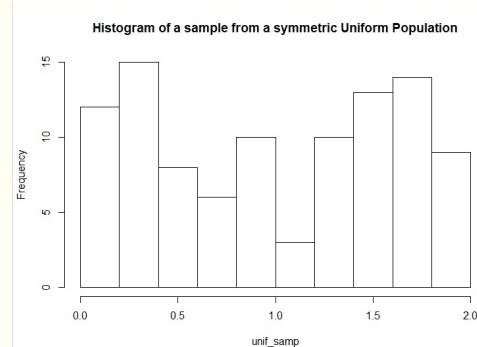
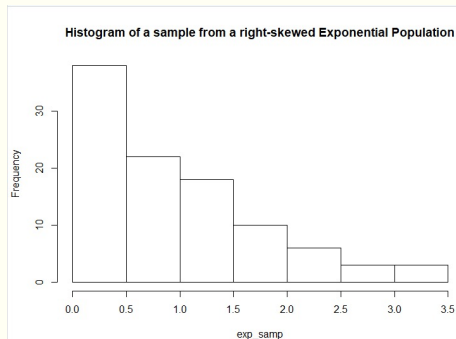
Checking the CLT - Simulation

- Let us use simulation to see if the Central Limit Theorem for \bar{X} is true.
- We will take samples of size 100 from an exponential population distribution (positively skewed or right-skewed), a uniform population distribution (symmetric) and a beta distribution with appropriate parameters (negatively skewed or left-skewed)
- The histogram of each sample is shown along with R code.

```
#  
exp_samp <- rexp(100, 1)      Generate 100 exponential (1) random variables  
mean(exp_samp)  
hist(exp_samp,  
      main=paste("Histogram of a sample from a right-skewed Exponential Population" ))  
..
```

```
#  
unif_samp <- runif(100, 0, 2) Generate 100 Uniform (0, 2) random variables  
mean(unif_samp)  
hist(unif_samp,  
      main=paste("Histogram of a sample from a symmetric Uniform Population" ))  
..
```

```
#  
beta_samp <- rbeta(100, 50, 1, ncp = 0)      Generate 100 beta (50, 1) random variables  
mean(beta_samp)  
hist(beta_samp,  
      main=paste("Histogram of a sample from a left-skewed Beta Population" ))  
..
```



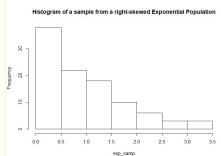
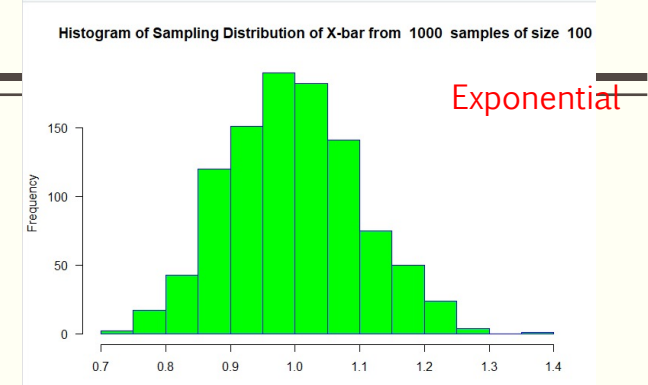
Testing the CLT - Simulation

- We now proceed to do the same simulation of the sampling distribution of \bar{X} that we did for the normal population earlier for each of the three distributions: exponential, uniform and beta, by taking 1000 samples, each of size 100.

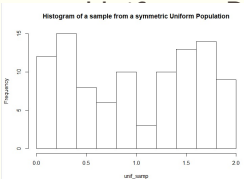
```
109 # Generating an empirical sampling distribution of sample mean - x-bar|
110 # Define the x_bar vector
111 num_samp = 1000
112 samp_size = 100
113 x_bar <- vector("numeric", num_samp)
114 #
115 # Each x-bar is the mean of a random sample of size 100 drawn from a 3 different distributions
116 # exponential (right-skewed), uniform (symmetric), beta (left-skewed)
117 # We are generating 1000 X-bars (from 1000 samples) and storing them in the x-bar vector
118 #
119 for (i in 1:num_samp) {
120   # Uncomment the distribution to be used in the next three lines; leave the other two commented
121   # x_bar[i] = mean(rexp(samp_size, 1))
122   # x_bar[i] = mean(runif(samp_size, 0, 2))
123   x_bar[i] = mean(rbeta(samp_size, 50, 1, ncp = 0))
124 }
125 #
126 # Calculate the mean and standard deviation (called standard error) of x-bar
127 # from the empirical sampling distribution formed by 1000 samples of size 50
128 #
129 Expec_x_bar <- mean(x_bar)
130 stderr <- sd(x_bar)
131 print(paste("The Expected value of X-bar is: ", round(Expec_x_bar,4)))
132 print(paste("The standard error or standard deviation of X-bar is: ", round(stderr,4)))
133 hist(x_bar,
134       main=paste("Histogram of Sampling Distribution of X-bar from ",num_samp,
135                 " samples of size ", samp_size, ""),
136       xlab="X-bar",
137       border="blue",
138       col="green",
139       # xlim=c(5, 15),
140       las=1,
141       breaks=20)
```

Testing the CLT - Simulation

- You can see from the simulation of the sampling distribution of \bar{X} that at sample size 100, regardless of the population distribution, the sampling distribution approaches normality.
- The mean of the sampling distribution of \bar{X} is close to the population mean and the standard deviation is the population standard deviation / $\sqrt{100}$ i.e., \sqrt{n} (n = 100)
- Exponential: Population mean = 1, population standard deviation = 1

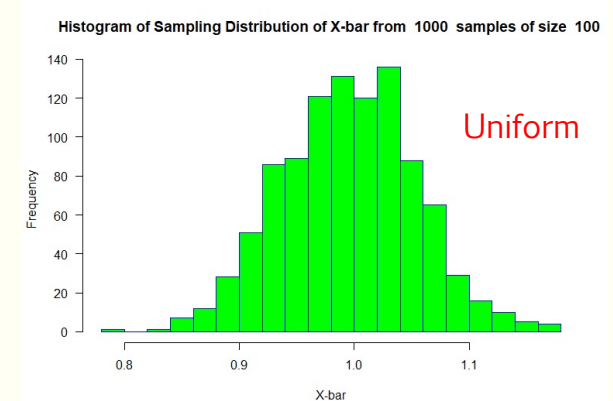


```
> print(paste("The Expected value of X-bar is: ", round(Expec_x_bar,4)))
[1] "The Expected value of X-bar is: 0.9976"
> print(paste("The standard error or standard deviation of X-bar is: ", round(stderr,4)))
[1] "The standard error or standard deviation of X-bar is: 0.1012"
```

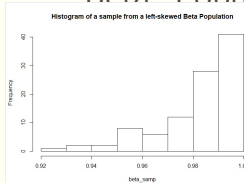


population mean = 1, population standard deviation = $\sqrt{4/12} = 0.5773$

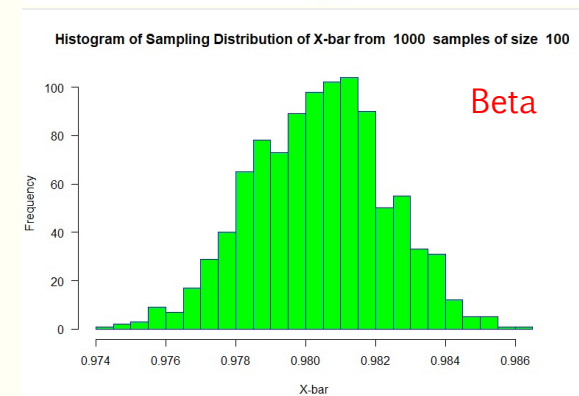
```
> print(paste("The Expected value of X-bar is: ", round(Expec_x_bar,4)))
[1] "The Expected value of X-bar is: 0.9952"
> print(paste("The standard error or standard deviation of X-bar is: ", round(stderr,4)))
[1] "The standard error or standard deviation of X-bar is: 0.0583"
```



- Beta: Population mean = $50/51 = 0.9804$, population standard deviation = 0.0192



```
> print(paste("The Expected value of X-bar is: ", round(Expec_x_bar,4)))
[1] "The Expected value of X-bar is: 0.9804"
> print(paste("The standard error or standard deviation of X-bar is: ", round(stderr,4)))
[1] "The standard error or standard deviation of X-bar is: 0.0019"
```



Case 3: Using the CLT – Book Example 7.8 – Page 406

- A study involving stress is conducted among the students on a college campus. **The stress scores follow a uniform distribution** with the lowest stress score equal to one and the highest equal to five. Using a sample of 75 students, find:
 - The probability that the **mean stress score** for the 75 students is less than two.
 - Want value will the **mean stress score** be for 90% of the samples? (This is the 90th percentile)

Solution:

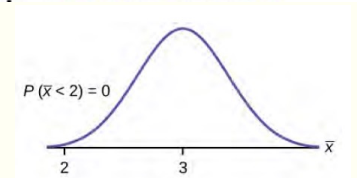
- In the population, $X = \text{Stress score} \sim \text{Uniform}(1, 5)$. Hence:

$$\mu = (5 + 1)/2 = 3$$

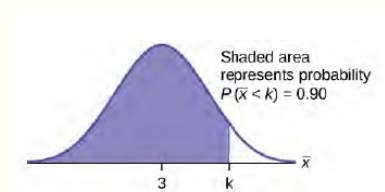
$$\sigma = \sqrt{(5 - 1)^2/12} = \sqrt{16/12} = 1.1547$$
- Since the population is not normal, $n = 75$ (greater than 30) and the population $\sigma = 1.1547$ is known, using the Central Limit Theorem $\bar{X} \sim N(3, 1.1547/\sqrt{75}) \sim N(3, 0.1333)$
 - probability that the **mean stress score** for the 75 students is less than two

$$= P(\bar{X} < 2) = P(Z < (2 - 3)/0.1333) = P(Z < -7.50) = 0$$
 - In 90% of the samples the **mean stress score** will be less than $P(\bar{X} < ?) = 0.9$. From z-tables, the z value is 1.282. Hence, $\bar{X} = 3 + 0.1333 \cdot 1.282 = 3.1708 =$

```
> pnorm(2, 3, 0.1333)
[1] 3.145556e-14
```



```
> qnorm(0.9, 3, 0.1333)
[1] 3.170831
```



Summary - Sampling Distribution of \bar{X}

- If the **population** distribution is assumed **normal**, the sampling distribution of \bar{X} is
 - Normal $(\mu, \frac{\sigma}{\sqrt{n}})$ if the population standard deviation σ is known. Use the standard normal **z** distribution, regardless of sample size.
 - Normal $(\mu, \frac{S}{\sqrt{n}})$ if the population standard deviation σ is **unknown** and the sample size is at least 30. Use the standard normal **z** distribution.
 - t-distribution with $(n-1)$ degrees of freedom, standard deviation σ is unknown and the sample size is small (< 30)
- If the **population** distribution is **not** assumed **normal**, and if the population standard deviation σ is known, the sampling distribution of \bar{X} , based on the Central Limit Theorem, is
 - Normal $(\mu, \frac{\sigma}{\sqrt{n}})$ and the sample size is at least 30. Use the standard normal **z** distribution.
- If the **population** distribution is **not** assumed **normal**, and if the population standard deviation σ is **not** known, one option to obtain the sampling distribution of \bar{X} is Bootstrapping.

Summary - Obtaining Sampling Distribution of \bar{X}

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z -distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{S}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{S}{\sqrt{n}})$ with $(n-1)$ degrees of freedom if Case 2 sample size < 30. Use t-distribution.
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size $\ll 30$, we really have to assume the population distribution. To avoid this, collect a larger sample.</i> Case 3	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)



PROBLEMS

Understanding and Using the Sampling Distribution of \bar{X}



Book Problem - Modified Version of 8.110 - page 483

- Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The Table below shows the ages of the corporate CEOs for a random sample of these firms. Assume the underlying population is normal.

48	58	51	61	56	59	74	63	53	50
59	60	60	57	46	55	63	57	47	55
57	43	61	62	49	67	67	55	55	49

```
> forbes_data <- c(48, 58, 51, 61, 56, 59, 74, 63, 53, 50, 59, 60, 60, 57, 46, 55, 63, 57, 47, 55, 57, 43, 61, 62, 49, 67, 67, 55, 55, 49)
> n <- length(forbes_data)
> print(n)
[1] 30
> print(paste("The sample x-bar is: ", round(mean(forbes_data),4)))
[1] "The sample x-bar is: 56.5667"
> print(paste("The sample standard deviation is: ", round(sd(forbes_data),4)))
[1] "The sample standard deviation is: 6.9067"
```

- What is the sampling distribution of \bar{X} = “sample mean age of CEOs of best small firms of 2012” ? Why?

Answer:

This is Case 2 with the population normal, population standard deviation unknown and sample size = 30. You can use either the z-distribution i.e., $\bar{X} \sim \text{Normal}(\mu, s/\sqrt{n})$

$$\bar{X} = 56.86, s = 6.83 \text{ and } \frac{s}{\sqrt{n}} = \frac{6.83}{\sqrt{30}} = 1.247; \text{ So, } \bar{X} \sim \text{Normal}(\mu, 1.247)$$

- What is our best estimate of μ using our sample (indicate units)?

Answer: $\bar{X} = 56.5667$ years

- What is our best estimate of σ using our sample (indicate units)?

Answer: $s = 6.9067$ years

- What is the expected value $E(\bar{X})$ (indicate units)?

Answer: μ in years.

- What is the standard error i.e., standard deviation of (\bar{X}) (indicate units)?
= 1.261 years

Answer: $\frac{s}{\sqrt{n}} = \frac{6.9067}{\sqrt{30}}$



Book Problem - Modified Version of 8.101 - page 481

- The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. For a random selection of 40 House candidates $\bar{X} = \$568,873$. The population standard deviation for this data to the nearest hundred is $\sigma = \$909,200$.

- What is the sampling distribution of \bar{X} = “sample mean of campaign contributions” ? Why?

Answer:

This is Case 3 with the population distribution unknown population standard deviation known. Sample size is 40.

So we use the Central Limit Theorem. $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ i.e., $\bar{X} \sim \text{Normal}(\mu, \frac{909200}{\sqrt{40}})$

- What is our best estimate of μ using our sample (indicate units)?

Answer: $\bar{X} = \$568,873$

- What is σ (indicate units)?
\$909,200

Answer: $\sigma =$

- What is the expected value $E(\bar{X})$ (indicate units)?

Answer: μ in dollars.

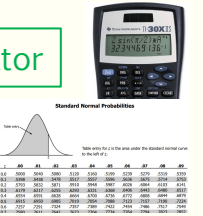
- What is the standard error i.e., standard deviation of (\bar{X}) (indicate units)?
43757.14

Answer: $\frac{909200}{\sqrt{40}} = \$$



Book Problem - Modified Version of 8.96 - page 479

- Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a (sample) standard deviation of 1.28 days. Assume we are interested in estimating the mean conference length of all engineering conferences (population mean μ).
- What is the sampling distribution of \bar{X} = “sample *mean* length of the conferences” ? Why?
 - **Answer:** We **don't know** the **population standard deviation**, so we have to assume that the population random variable X = “length of conferences” is normally distributed with unknown population mean μ (population *mean* length of the conferences) and unknown standard deviation σ .
 - $\bar{X} \sim N(\mu, \frac{s}{\sqrt{n}}) \sim N(\mu, \frac{1.28}{\sqrt{84}}) \sim N(\mu, 0.1397)$
- What is our best estimate of μ using our sample (indicate units)? **Answer:** $\bar{X} = 3.94$ days
- What is our best estimate of σ using our sample (indicate units)? **Answer:** $s = 1.28$ days
- What is the expected value $E(\bar{X})$ (indicate units)? **Answer:** μ in days
- What is the standard error i.e., standard deviation of (\bar{X}) (indicate units)? **Answer:** 0.1397 days



Book Problem - Modified Version of 8.96 - page 479 (continued)

- **Suppose** that the true population mean for duration of engineering conferences $\mu = 4$ days.
 - Find the probability that the sample mean of another sample (of the same size) is greater than your particular sample mean of 3.94 days.
 - Answer: $\bar{X} \sim N(4, 0.1397)$

$$\text{So } P(\bar{X} > 3.94) = P(z > ((3.94 - 4)/0.1397)) = P(z > -0.4249) = 0.665$$
 - What percentage of the sample means of all possible samples of the same size *will be less* than the sample mean of our sample?
 - Answer: $\bar{X} \sim N(4, 0.1397)$

$$\text{So } P(\bar{X} > 3.94) = 1 - P(z < -0.4249) = 33.5\%$$
 - What percentage of the sample means of all possible samples of the same size will be within 2 standard deviations of $\mu = 4$? Will it work for any value of μ ?
 - Answer: $\bar{X} \sim N(4, 0.1397)$

$$\text{So } P(3.72 < \bar{X} < 4.28) = P(((3.72 - 4)/0.1397) < z < ((4.28 - 4)/0.1397))$$

$$= P(-2 < z < 2) = 0.46 \text{ so } 4.6\%.$$
 - Percentage within 2 standard deviations will be $1 - 0.046 = 95.4\%$
 - Let $\mu = 5$. So $P(4.72 < \bar{X} < 5.28) = P(-2 < z < 2) = 0.46$ so 4.6%. So does not depend on μ .



EFFECT OF SAMPLE SIZE AND STANDARD ERROR

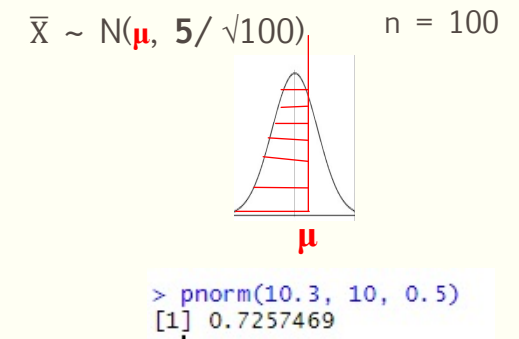
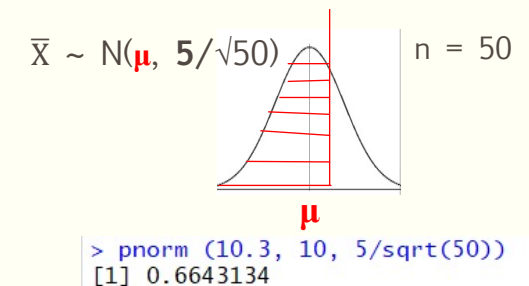
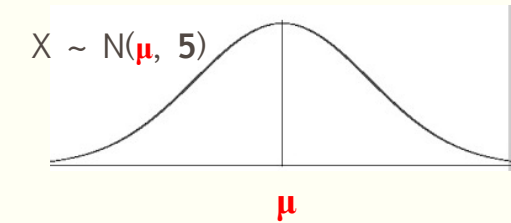
Standard Normal Probabilities

Area under the standard normal curve to the left of Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5518	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7122	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7421	0.7453	0.7484	0.7515	0.7546
0.7	0.7577	0.7607	0.7637	0.7667	0.7696	0.7725	0.7754	0.7783	0.7811	0.7839
0.8	0.7867	0.7895	0.7923	0.7950	0.7977	0.8004	0.8031	0.8058	0.8085	0.8112
0.9	0.8139	0.8166	0.8192	0.8218	0.8244	0.8269	0.8294	0.8319	0.8344	0.8369
1.0	0.8394	0.8413	0.8438	0.8463	0.8488	0.8512	0.8536	0.8559	0.8582	0.8605
1.1	0.8628	0.8649	0.8670	0.8691	0.8712	0.8732	0.8752	0.8771	0.8790	0.8810
1.2	0.8829	0.8848	0.8867	0.8886	0.8905	0.8923	0.8941	0.8959	0.8977	0.8995
1.3	0.9012	0.9029	0.9046	0.9062	0.9079	0.9095	0.9111	0.9127	0.9143	0.9158
1.4	0.9174	0.9189	0.9205	0.9221	0.9236	0.9251	0.9266	0.9281	0.9296	0.9311
1.5	0.9324	0.9340	0.9354	0.9369	0.9384	0.9398	0.9413	0.9427	0.9441	0.9455
1.6	0.9469	0.9483	0.9496	0.9510	0.9524	0.9538	0.9552	0.9565	0.9579	0.9592
1.7	0.9606	0.9619	0.9632	0.9645	0.9658	0.9671	0.9684	0.9696	0.9709	0.9721
1.8	0.9732	0.9744	0.9756	0.9768	0.9779	0.9790	0.9801	0.9812	0.9823	0.9834
1.9	0.9844	0.9854	0.9864	0.9874	0.9884	0.9894	0.9904	0.9913	0.9922	0.9931
2.0	0.9940	0.9949	0.9957	0.9965	0.9973	0.9980	0.9987	0.9994	0.9999	1.0000

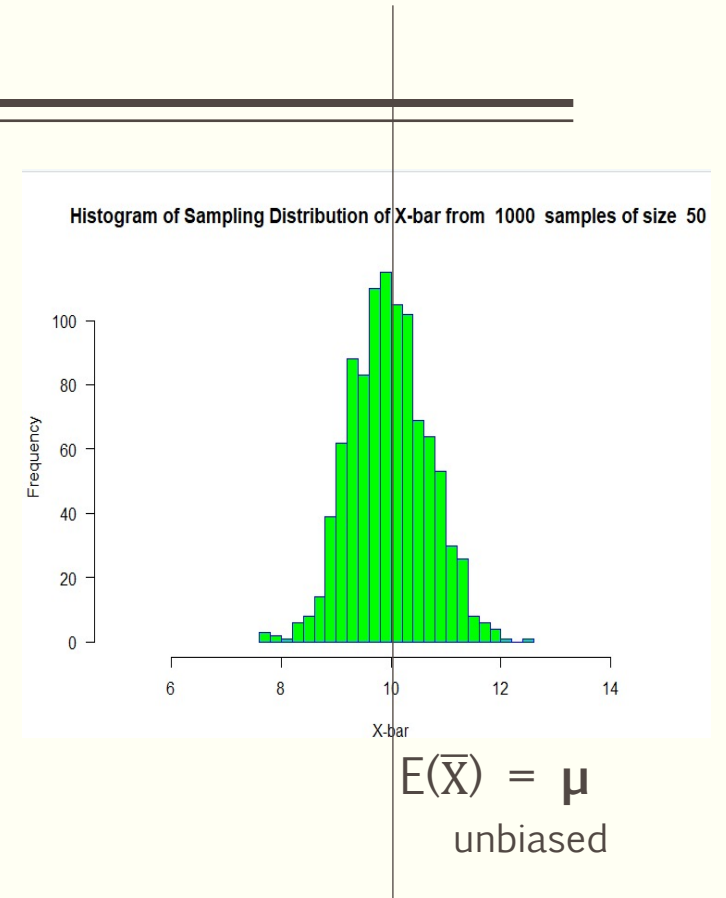
Effect of Sample Size and Standard Error

- When we increase the sample size (n), the standard Error $\frac{\sigma}{\sqrt{n}}$ will reduce.
- This means that the sampling distribution of \bar{X} will have a smaller dispersion from the mean μ (smaller standard deviation). That is, *there is less variation in \bar{X} from sample to sample.*
- Two things happen:
 - Sample means (\bar{X}) will be closer to each other compared to a smaller sample size
 - Each sample mean (\bar{X}) will be closer to the unknown μ , compared to a smaller sample size
- Example:
 - Suppose I take a sample of size 100 from a population that is assumed normal $N(\mu, 5)$
 - If my sample mean $\bar{X} = 10.3$, how likely was the sample to have come from a population with $\mu = 10$?
 - Sampling distribution of $\bar{X} \sim N(\mu, 5/\sqrt{n})$ or $\sim N(\mu, 0.5)$ since $n = 100$.
 - Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 10) = P(Z \leq \frac{10.3 - 10}{0.5}) = 0.726$ (for $n = 100$)
 - Probability $P(\bar{X} \leq 10.3 \text{ assuming } \mu = 10) = P(Z \leq \frac{10.3 - 10}{0.7071}) = 0.664$ (for $n = 50$)
 - When sample size increases the probability that our sample (mean) came from a population with $\mu = 10$ is greater..



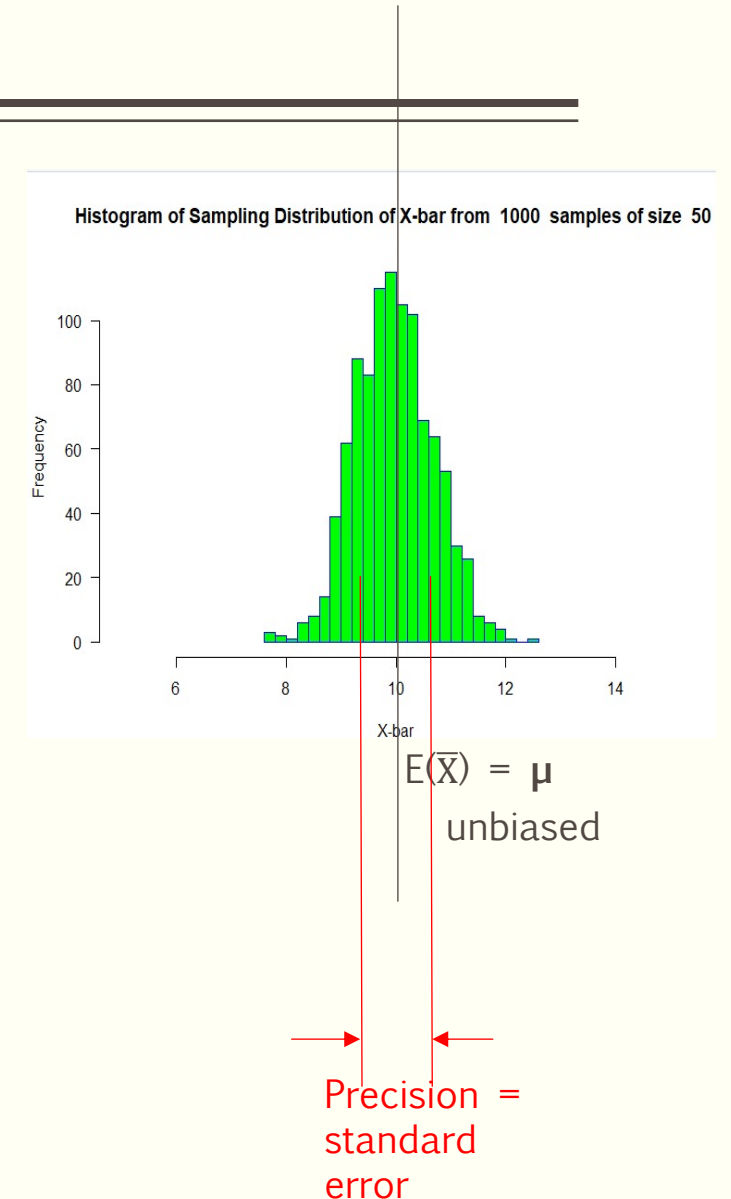
Evaluating Estimators – Bias and Precision

- We saw that we can use the sample mean \bar{X} to estimate the population parameter μ .
- Regardless of the population distribution, it is always true that the expected value of the sampling distribution of $\bar{X} = \mu$.
- That is, from all possible samples of a certain size, $E(\bar{X}) = \mu$.
- How do we understand this Expected value? If we average the \bar{X} from all possible samples, this average will be exactly the population mean.
- Thus, \bar{X} is said to be an *unbiased estimator* of μ . This is a *desirable property* of any estimator.
- Bias is the difference between the Expected Value of Sampling Distribution and the population parameter.
- Not all estimators are unbiased (such as if you use the sample median to estimate μ). That is, there will be a difference between the Expected Value of Sampling Distribution and the population parameter. In this case, the estimator is said to be *biased*.



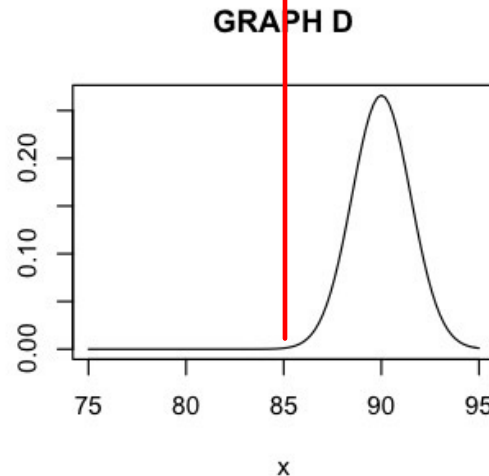
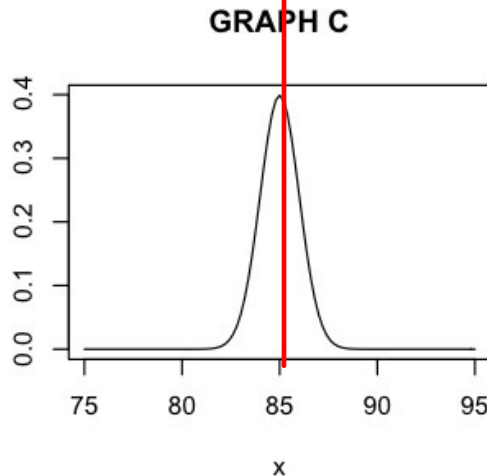
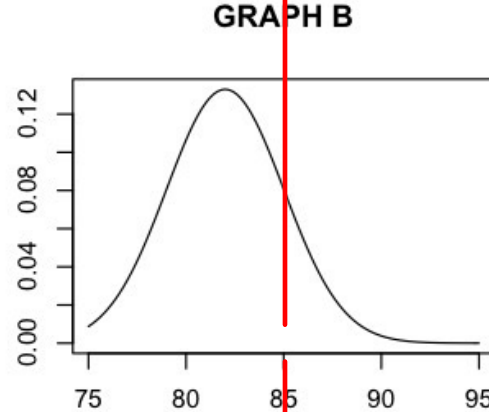
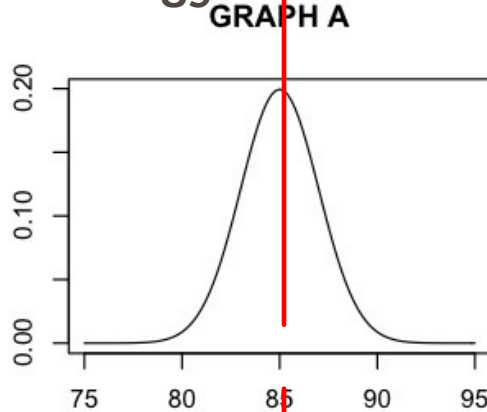
Evaluating Estimators – Bias and Precision

- The *precision* of an estimator is how close the estimator is to the true population parameter, from sample to sample.
- In the case of \bar{X} , it is how close \bar{X} is to μ , from sample to sample.
- Thus, it can be measured by the spread of \bar{X} values in the sampling distribution.
- It is usually measured by the standard deviation of the sampling distribution, or the standard error (depending on the case, for \bar{X} , the standard error is either $\frac{\sigma}{\sqrt{n}}$ or $\frac{s}{\sqrt{n}}$.
- We want precision to be high (i.e., *the standard deviation or standard error to be low*), because then the \bar{X} from any sample is close to the true population μ . If the standard error is large, then the sample information is less reliable.
- The *easiest way to increase precision is to increase the sample size*, because *standard error always goes down with increasing sample size*.



Evaluating Estimators – Bias and Precision

True population mean =
85



- Assume unknown population **parameter value 85**.
- The figure shows 4 different sampling distributions (different statistics or estimators were used)
- Which is the most desirable?
- A & C are unbiased, D & B are biased
 - C has no bias, and highest precision (smallest SE)
 - D has bias, but second highest precision (lowest SE)
 - A has no bias, and third highest precision (lowest SE)
 - B is the worst estimator because it has bias and lowest precision (highest SE)



CONFIDENCE INTERVALS

Book Chapter 8

Using the Sampling Distribution of \bar{X} to Estimate μ

- Now that we know what the sampling distribution of \bar{X} is and how to obtain it (at least for some of the common situations), how do we use it?
- The key is to remember the following:
 - We collect only one sample (which will give us \bar{X} , (the sample mean of the single sample), and s (the standard deviation of single sample), both of which are statistics because they are computed from the sample.
 - $E(\bar{X})$ (the mean of the sampling distribution of \bar{X}) = μ (i.e., \bar{X} is an *unbiased* estimator of the unknown population parameter μ we are trying to estimate)
 - $SD(\bar{X})$ (the standard deviation of the sampling distribution of \bar{X}) = Standard Error = σ/\sqrt{n} or s/\sqrt{n} is a measure of *precision* of our estimator \bar{X} in estimating μ .
 - The sampling distribution of \bar{X} is either Normal or t-distribution
- Using these facts, we can see how good our single sample estimate is i.e., the quality of our single sample \bar{X} .

Confidence Intervals - How Good is our Single Sample \bar{X} in Estimating μ ?

- Suppose the population mean IQ is an unknown (μ) with a standard deviation of (σ)=20. We take a single sample of size 100, and find that the sample mean \bar{X} is 105. We don't know the population distribution.
- The **CLT** says that the sampling distribution of \bar{X} will be Normally distributed ($\mu, \frac{\sigma}{\sqrt{n}}$) i.e., the sampling distribution of the sample mean will have the distribution. $N(\mu, 2)$.
- What is our best guess *point estimate* of μ ?
 - Answer: We only have one estimate, namely \bar{X} , so our answer is 105.
- How *precise* is our answer?
 - Answer: We will give a *confidence interval* for μ
- Confidence Intervals (or more precisely, $(1-\alpha)\%$ Confidence Intervals) are called *interval estimates of the population parameter*:
 - 90% CI = $(1 - 0.1) \cdot 100\%$ CI implies $\alpha = 0.1$.
 - 95% CI = $(1 - 0.05) \cdot 100\%$ CI implies $\alpha = 0.05$.
 - 99% CI = $(1 - 0.01) \cdot 100\%$ CI implies $\alpha = 0.01$.

Confidence Interval for μ

- Since we know the sampling distribution of \bar{X} is $N(\mu, \frac{\sigma}{\sqrt{n}})$ we can use this information to obtain an interval that contains μ , with level of confidence $(1 - \alpha)\%$ where α is between 0 and 1.
 - For example, $\alpha = 0.05$ means $(1 - \alpha) = 0.95$, so we will have a 95% confidence interval (95% CI)

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$Z = (\bar{X} - \mu) / (\frac{\sigma}{\sqrt{n}}) \sim N(0,1)$$

- From the diagram, $z_{\alpha/2}$ is **the number of standard errors \bar{X} is away from μ .**

- Probability** $(\mu - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}) = (1 - \alpha)$.

- Remember, \bar{X} is a random variable and has a probability.

- $(\mu - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$ can also be written as $(\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$

- But μ is NOT A RANDOM VARIABLE. It is an UNKNOWN **CONSTANT**.

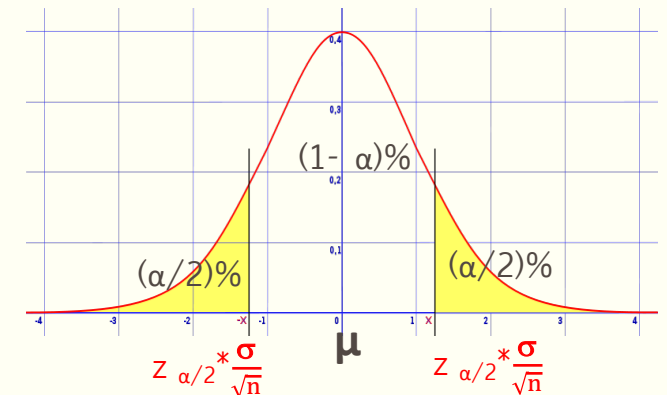
- So we cannot use the term “Probability”; i.e., we **CANNOT** say

Probability $(\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}) = (1 - \alpha)$.

- Instead, we use the term **confidence interval**.

$(1 - \alpha)\%$ Confidence Interval for $\mu = (\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$

- i.e., $(\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$ is a $(1 - \alpha)\%$ Confidence Interval. The term “ $z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$ ” is called the **error bound for the mean or EBM**



90% CI = $(1 - 0.1) * 100\%$ CI implies $\alpha = 0.1$.

95% CI = $(1 - 0.05) * 100\%$ CI implies $\alpha = 0.05$.

99% CI = $(1 - 0.01) * 100\%$ CI implies $\alpha = 0.01$.

How Good is our single sample \bar{X} in Estimating μ ?

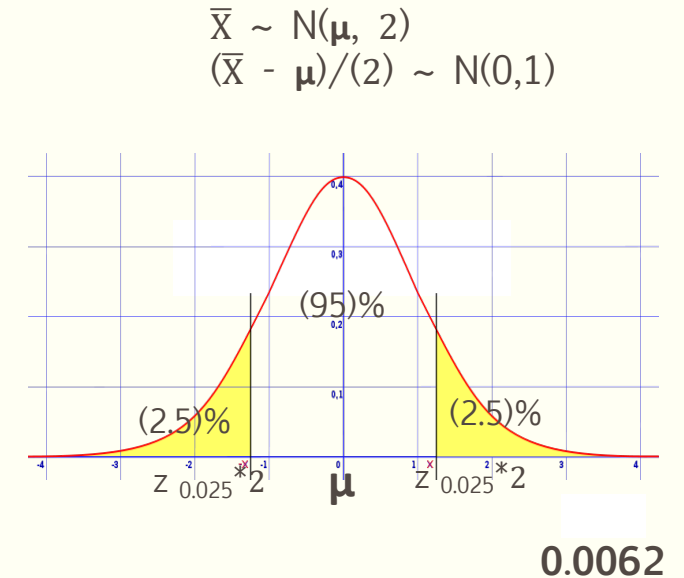
Practice with Tables

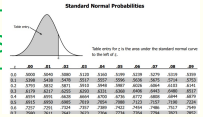
Standard Normal Probabilities

Area under the standard normal curve to the left of Z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7122	.7156	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7421	.7453	.7484	.7515	.7546
0.7	.7577	.7607	.7637	.7667	.7696	.7725	.7754	.7783	.7811	.7839
0.8	.7867	.7895	.7923	.7950	.7977	.8004	.8031	.8058	.8085	.8112
0.9	.8139	.8166	.8192	.8218	.8244	.8269	.8294	.8319	.8344	.8369
1.0	.8394	.8419	.8443	.8468	.8491	.8515	.8538	.8561	.8584	.8607
1.1	.8629	.8651	.8673	.8694	.8715	.8735	.8755	.8774	.8793	.8812
1.2	.8830	.8849	.8868	.8887	.8905	.8923	.8940	.8957	.8974	.8992
1.3	.9009	.9026	.9042	.9059	.9075	.9092	.9108	.9124	.9141	.9157
1.4	.9172	.9188	.9204	.9219	.9235	.9251	.9266	.9281	.9296	.9312
1.5	.9327	.9342	.9357	.9372	.9387	.9401	.9416	.9429	.9443	.9457
1.6	.9471	.9484	.9497	.9511	.9524	.9537	.9550	.9562	.9575	.9588
1.7	.9599	.9611	.9623	.9635	.9646	.9657	.9668	.9678	.9688	.9698
1.8	.9709	.9719	.9729	.9738	.9747	.9756	.9765	.9773	.9782	.9790
1.9	.9799	.9808	.9816	.9824	.9832	.9840	.9848	.9856	.9864	.9871
2.0	.9879	.9886	.9893	.9900	.9906	.9913	.9919	.9925	.9931	.9936
2.1	.9941	.9946	.9951	.9955	.9959	.9964	.9968	.9972	.9976	.9980
2.2	.9984	.9988	.9991	.9994	.9996	.9998	.9999			
2.3										
2.4										
2.5										
2.6										
2.7										
2.8										
2.9										
3.0										

- For our Example:
 - $\bar{X} \sim N(\mu, 2)$
- 95% CI:
 - 95% CI means $(1 - \alpha) = 0.95$ or $1 - 0.95 = \alpha = 0.05$
- At $\alpha = 0.05$, from Z-table, $z_{0.025} = 1.96$; \bar{X} is 1.96 standard errors away from μ .
- Error Bound for Mean = EBM = $z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = 1.96 * 2 = 3.92$
- Hence, our 95% CI for $\mu = (\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$
 - $= (105 - 3.92, 105 + 3.92)$
 - $= (101.08, 108.92)$
- What does this 95% *Confidence* Interval say, if it is **not** a probability?
- It says, “In 95% of the samples of size n, where the confidence Interval is constructed as $(\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$, the unknown population mean μ can be found in the confidence interval; only 5% of sample confidence intervals will not contain μ .”





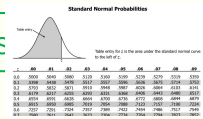
What does the Confidence Interval Mean (and NOT Mean)?

- The Confidence Interval does **not** mean the following:
 - Probability (that the unknown population mean μ is between 101.08 and 108.92) = 0.95
 - We are 95% “confident” that the unknown population mean μ is between 101.08 and 108.92
- Why?
 - Because μ is NOT A RANDOM VARIABLE (we are using frequentist, not Bayesian statistics). It is an UNKNOWN **CONSTANT** and a single number. It cannot be “between” two values and has no probability for being between two values.
- So what **does** it mean?
 - To answer this, ask yourself what a sampling distribution is. It represents the distribution of all possible (infinite) samples of the same size.
 - If I take another sample, the \bar{X} will be different and therefore 95% CI = $(\bar{X} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}})$ will be different.
 - However, the **EBM**, $z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$ **will be the same** for a particular α , in this case 1.96.
 - So let us consider 10 samples and look at the 95% CI.

What does the Confidence Interval Mean (and NOT Mean)?

Sample	\bar{X}	95% CI
1	105	(101.08, 108.92)
2	100	(96.08, 103.92)
3	80	(76.08, 83.92)
4	120	(116.08, 123.92)
5	102	(98.08, 105.92)
6	104	(100.08, 107.92)
7	105	(101.08, 108.92)
8	101	(97.08, 104.92)
9	95	(91.08, 98.92)
10	90	(86.08, 93.92)

- So what **does** CI mean?
 - If we consider 10 samples and look at the 95% CI what we can talk about is whether a particular CI **contains** μ (the fixed population mean) **or not**.
 - The 95% confidence interval tells us that *if we keep taking samples infinitely*, and constructing intervals based on $(\bar{X} \pm \text{EBM})$ for each sample, **then 95% of these intervals will contain the true population mean μ** .
 - What good is this? One use is to separate plausible values of μ from implausible ones.
 - Suppose you think μ should be 80. Then the 95% CI from our sample, namely (101.08, 108.92) tells us that it is implausible to the extent that we expect 95% of the intervals to not contain this value.
 - On the other hand, if you think μ should be 103, then based on the (101.08, 108.92) 95% CI, we say that we expect 95% of intervals to contain this value, ~~so~~ ³² it is plausible in that sense.



What does the Confidence Interval Mean (and NOT Mean)?

- Suppose we feel that our statement of plausibility for the value of μ is not precise enough. We can express this in three ways:
 - a) We want a higher confidence, let us say 99% CI, for the same sample size, or
 - b) We want 95% confidence, but a much sharper interval (i.e., narrower EBM), or
 - c) We want both of above.
- a) 99% CI for same sample size:
 - 99% CI means $(1 - \alpha) = 0.99$ or $1 - 0.99 = \alpha = 0.01$; At $\alpha = 0.01$, from Z-table, $z_{0.005} = 2.575$
 - $EBM = 2 * 2.575 = 5.15$, so 99% CI (99.85, 110.15). The interval is wider, so more of the intervals (99% of them) are expected to contain the true population mean, μ .
- b) 95% CI for larger sample size (say 400):
 - 95% CI means $(1 - \alpha) = 0.95$ or $1 - 0.95 = \alpha = 0.05$; At $\alpha = 0.05$, from Z-table, $z_{0.025} = 1.96$
 - Now, $\sigma/\sqrt{n} = 20/20 = 1$ so $EBM = 1.96$, so 95% CI (103.04, 106.96). The interval is narrower, so we gained more precision (through smaller standard error) for intervals expected to contain the true population mean, μ .
- c) What sample size do we need to get a 99%CI with $EBM = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = 2.575 * \frac{20}{\sqrt{n}} = 1.96$
 - a) $\text{Sqrt}(n) = 2.575 * 20 / 1.96 = 26.3$ so $n = 692$.

Table of Useful Values When Sampling Distribution is Normal

$(1-\alpha)\%$ CI	α	$Z_{\alpha/2}$
90%	0.1	1.645
95%	0.05	1.96
99%	0.01	2.575

Confidence Interval

Practice with Tables

Example:

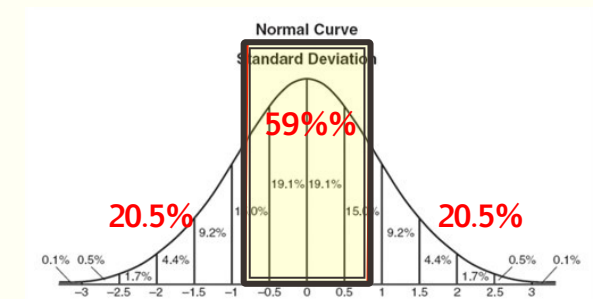
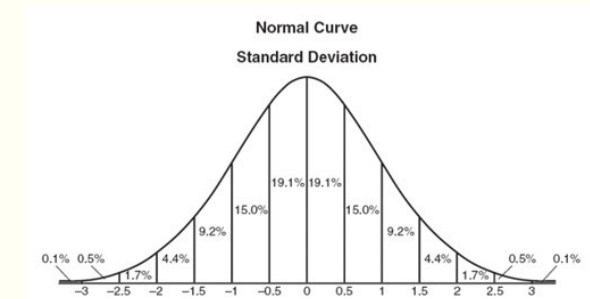
- The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The Bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.
 - Construct a 90% confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.
 - If the Census wants to increase its level of confidence to 98% and keep the error bound the same by taking another survey, what changes should it make?
 - If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with (n-1) degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size < 30, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)

Solution:

The sampling distribution of $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$.

- To get a 90% CI, there will be 5% on each side of the sample mean from a standard normal distribution. So $\alpha = 0.10$. Hence, $z_{\alpha/2} = 1.65$.
 Hence the 90% CI = $(\bar{X} - 1.65 * \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.65 * \frac{\sigma}{\sqrt{n}}) = (8.2 - 1.65 * \frac{2.2}{\sqrt{200}}, 8.2 + 1.65 * \frac{2.2}{\sqrt{200}}) = (8.2 - 0.256, 8.2 + 0.256) = (7.944, 8.456)$. The error bound is ± 0.256 .
 This interval says that if all possible samples were taken, 90% of the intervals constructed this way would contain the true average time taken to complete the short form of the US Census.
- When the confidence increases, $z_{\alpha/2}$ increases, σ remains the same. So to keep the error bound the same, n has to increase.
 For 98% CI, there will be 1% on each side of the sample mean. The $z_{\alpha/2} = \pm 2.326$. Hence, $2.326 * \frac{2.2}{\sqrt{n}} = 0.256$. Solving for n = $(2.326 * 2.2 / 0.256)^2$ we get n = 400. So the sample size has to be 400.
- If n = 50, $z_{\alpha/2} * \frac{2.2}{\sqrt{50}} = 0.256$. So $z_{\alpha/2} = 0.256 * \sqrt{50} / 2.2 = 0.8228$.
 Hence, the probability on one half of the CI is $1 - 0.795 = 0.205$ or 20.5%. So $\alpha = 0.41$.
 So we have an approximately 59% CI (see figure)

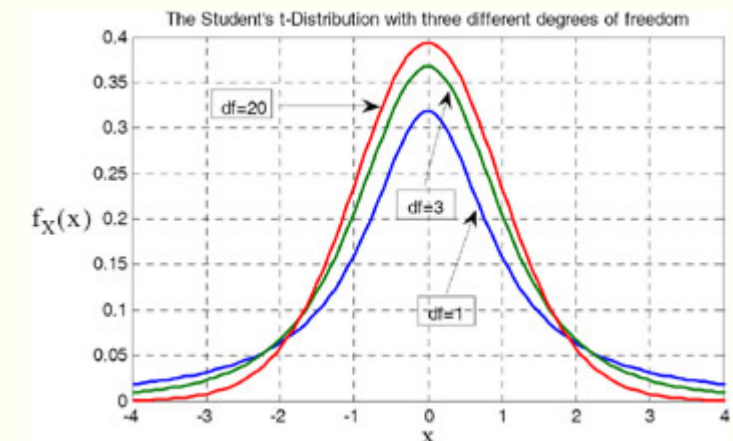


Confidence Interval Using The Student t-distribution

- The student t-distribution is used when we know the population is normal, but we don't know the population standard deviation (and the sample size is small, though at larger sample sizes it simply becomes the z-distribution)
- The Student t-distribution has a single parameter k called the *degrees of freedom*
- To get the confidence interval using the *t-score* (instead of the *z-value* or *z-score*)
 - Calculate the sample average, \bar{X} and the standard error SE of the sample mean $SE(\bar{X}) = \frac{s}{\sqrt{n}}$ where s is the standard deviation of the sample itself given by $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{(n-1)}}$, and n is the sample size
 - The degrees of freedom for the t-distribution is $n-1$
 - Construct $(\bar{X} - t_{\alpha/2} * \frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2} * \frac{s}{\sqrt{n}})$
- All the interpretations remain the same. The **error bound** is given by $t_{\alpha/2} * \frac{s}{\sqrt{n}}$
- In **R**, we can use the **qt(prob, df)** function to get $-t_{\alpha/2}$ and $t_{\alpha/2}$

```
> print(qt(0.025,29))
[1] -2.04523
> print(qt(0.975,29))
[1] 2.04523
```
- When the sample size is large, the degrees of freedom is large, and the t-score becomes very close to the z-score. So, for large sample sizes we can use the standard normal distribution

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with $(n-1)$ degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution If sample size $\ll 30$, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)



Using the Student t-distribution

■ Example:

- Consider the data set below as a sample. Calculate the 95% and 99% CI for the population mean of the Income variable assuming that it is normally distributed.

- Sample mean = \bar{X} = 165,790;
- The sample standard deviation s == 115348.25.
- $n=10$, so $\frac{s}{\sqrt{n}}$ = 36476.32.
- The degrees of freedom $k = n-1 = 9$

```
> print(qt(0.975,9))
[1] 2.262157
> print(qt(0.025,9))
[1] -2.262157
```

- 95% CI, there will be 2.5% on each side of the sample mean from a standard normal distribution. So $\alpha = 0.05$.

- $t_{\alpha/2}$ with 9 degrees of freedom and $(1 - \alpha/2) = 97.5\%$ cdf value is = 2.262.
- Error bound** is $+ t_{\alpha/2} * \frac{s}{\sqrt{n}} = 2.262 * 36476.32 = \pm 82509.44$.
- The 95% CI for Income = $165,790 \pm 82509.44$

```
> print(qt(0.995,9))
[1] 3.249836
> print(qt(0.005,9))
[1] -3.249836
```

- 99% CI, $\alpha = 0.01$ and $\alpha/2 = 0.005$

- $t_{\alpha/2}$ with 9 degrees of freedom and $(1 - \alpha/2) = 99.5\%$ cdf value is = 3.25.
- Error Bound** = $3.25 * 36476.32$
- 99% CI = $165,790 \pm 118,548.04$

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with $(n-1)$ degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size $<< 30$, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)

Name	Income	Net Worth	Sales
Adams, John	38,900	65,924	1,535
Ramesh, Jyoti	172,000	178,154	2,196
Mendez, Nick	218,000	265,209	1,287
Mendez, Joan	182,000	85,277	2,143
Ritter, Jake	434,000	193,760	707
Rao, Eric	82,000	314,953	2,170
Blake, Ann	112,000	192,946	1,229
Bishop, Marge	242,000	339,705	520
Ahmed, Mo	111,000	185,767	2,326
Shultz, Dante	66,000	97,778	588

Using the Central Limit Theorem

Practice with Tables

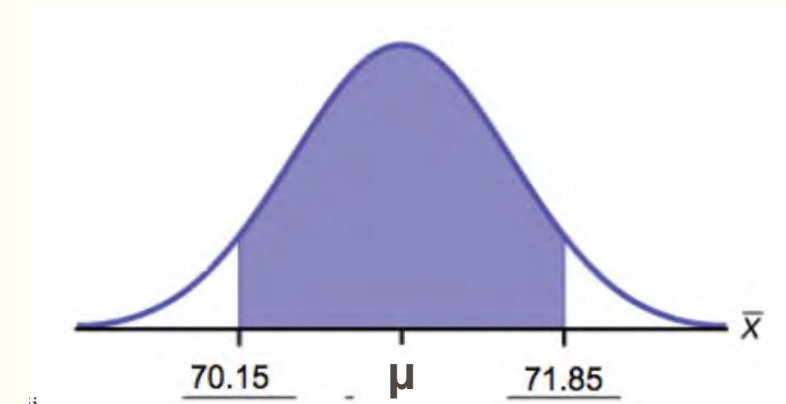
Standard Normal Probabilities

Table giving the area under the standard normal curve to the left of Z.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7122	.7156	.7189	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7421	.7453	.7484	.7515	.7546
0.7	.7577	.7608	.7638	.7668	.7697	.7726	.7755	.7784	.7812	.7841
0.8	.7869	.7896	.7924	.7952	.7979	.8006	.8033	.8060	.8086	.8113
0.9	.8139	.8166	.8192	.8218	.8244	.8269	.8294	.8319	.8344	.8369
1.0	.8394	.8419	.8444	.8468	.8493	.8517	.8541	.8564	.8588	.8613
1.1	.8636	.8659	.8681	.8704	.8726	.8748	.8769	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8906	.8925	.8943	.8961	.8978	.8995	.9012
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9237	.9251	.9266	.9281	.9296	.9310	.9324
1.5	.9340	.9354	.9368	.9382	.9396	.9409	.9423	.9437	.9450	.9464
1.6	.9477	.9490	.9504	.9517	.9529	.9542	.9554	.9566	.9578	.9590
1.7	.9601	.9613	.9625	.9636	.9647	.9658	.9669	.9679	.9689	.9699
1.8	.9709	.9719	.9729	.9738	.9748	.9757	.9766	.9775	.9784	.9793
1.9	.9801	.9810	.9819	.9827	.9836	.9844	.9852	.9860	.9868	.9876
2.0	.9884	.9892	.9900	.9907	.9915	.9922	.9929	.9936	.9943	.9950
2.1	.9956	.9963	.9969	.9975	.9980	.9985	.9990	.9994	.9998	.9999
2.2	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.3	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.4	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.5	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.6	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.7	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.8	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
2.9	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.0	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

- Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. Forty-eight male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.
- Construct a 95% confidence interval for the population mean height of male Swedes.
- Here we know the population standard deviation but do not know the population distribution. The sample size is 48, so we can use the Central Limit Theorem for the sampling distribution of \bar{X} i.e., $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$.
- $$\text{EBM} = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = 1.96 * 3 / \sqrt{48} = 0.849 \text{ inches.}$$
- 95% CI = (70.151, 71.849)
- 95% of the intervals (from infinite samples) constructed this way will contain the true population height of male Swedes in inches.

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with (n-1) degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size < 30, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)



Estimating Population Proportions

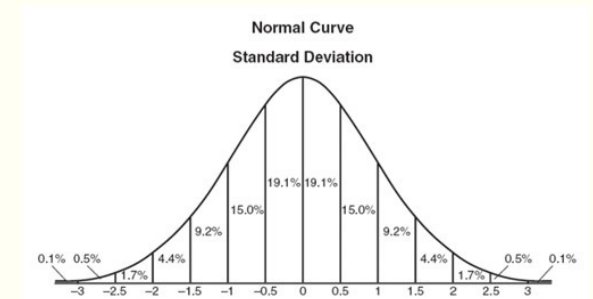
- Sometimes, instead of estimating the population mean μ , we want to estimate the population *proportion*. For example, if we consider our data set as a sample, we are interested in estimating the proportion of our population of customers who are female.
- Earlier, we saw that we can use the binomial distribution to calculate the probability of having the 4 females in our data set of size 10.
- In estimation, we reverse this problem, and ask ourselves: “*Given that we have 4 female customers in our sample of 10 records, what is the 95% confidence interval estimate for the proportion of females in our customer population*”.
- Let p represent the (unknown) proportion of females in the population and p' represent the same proportion in a sample.
- Then, using the normal approximation to the binomial, we have:
- The sampling distribution $p' \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
- Because we don't know the **population proportion** p , we use the sample proportion p' to get the standard error.
- That is, $p' \sim N(p, \sqrt{\frac{p'(1-p')}{n}})$; The error bound for a $(1 - \alpha)\%$ CI is $z_{\alpha/2} * \sqrt{\frac{p'(1-p')}{n}}$

ID	Name	Age	Gender
001	Adams, John	36	M
002	Ramesh, Jyoti	23	F
003	Mendez, Nick	67	M
004	Mendez, Joan	38	F
005	Ritter, Jake	24	M
006	Rao, Eric	61	M
007	Blake, Ann	26	F
008	Bishop, Marge	44	F
009	Ahmed, Mo	31	M
010	Shultz, Dante	44	M

Estimating Population Proportions

- $p' = x / n$ where x represents the number of successes and n represents the sample size. The variable p' is the sample proportion and serves as the **point estimate** for the true population proportion p of successes (much like the sample mean is the point estimate of the population mean). In our example, $p' = x / n = 4/10 = 0.4$.
- $p' \sim N(p, \sqrt{\frac{p'(1-p')}{n}})$; The error bound for a $(1 - \alpha)\%$ CI is $z_{\alpha/2} * \sqrt{\frac{p'(1-p')}{n}}$
- Therefore, for our data set the **error bound** for a 95% CI is (given $z_{\alpha/2} = 1.96$)
 $= 1.96 * \sqrt{\frac{0.4*0.6}{10}} = 0.30$
- The 95% CI for the proportion of female customers in our population is (0.1, 0.7).
- Notice that this is too wide to be of use. Why? Because the standard error is too high. What should you do to get an error bound of 0.05 for a 95% CI? We need to increase the sample size.
- Sample size needed to get error bound of 0.05: $1.96 * \sqrt{((0.4*0.6)/n)} = 0.05$ so $\sqrt{n} = (1.96 * \sqrt{0.24})/0.05 = 19.2$. So $n = 369$.

ID	Name	Age	Gender
001	Adams, John	36	M
002	Ramesh, Jyoti	23	F
003	Mendez, Nick	67	M
004	Mendez, Joan	38	F
005	Ritter, Jake	24	M
006	Rao, Eric	61	M
007	Blake, Ann	26	F
008	Bishop, Marge	44	F
009	Ahmed, Mo	31	M
010	Shultz, Dante	44	M



“Plus Four” Confidence Interval for p (Book Page 460)

- There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.
- Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals. We simply pretend that we have four additional observations. Two of these observations are successes and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.
- Computer studies have demonstrated the effectiveness of this method. It should be used when the confidence level desired is at least 90% and the sample size is at least ten.

“Plus Four” Confidence Interval for p (Book Page 460)

- For our example of proportion of females:
- $p' = x / n = 4/10 = 0.4$. Using “Plus Four” we have $x+2/n+4 = 6/14 = 0.429$
- $p' \sim N(p, \sqrt{\frac{p'(1-p')}{n}})$; The error bound for a $(1 - \alpha)\%$ CI is $z_{\alpha/2} * \sqrt{\frac{p'(1-p')}{n}}$
- Therefore, for our data set the **error bound** for a 95% CI is (given $z_{\alpha/2} = 1.96$) $= 1.96 * \sqrt{\frac{0.429 * 0.571}{14}} = 0.259$
- The 95% CI for the proportion of female customers in our population is (0.17, 0.688).



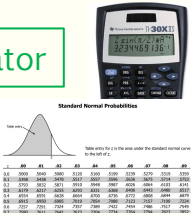
PROBLEMS

Confidence Intervals

Book Problem - 8.96 - page 479

Practice with calculator

Practice with Tables



- Announcements for 84 upcoming engineering conferences were randomly picked from a stack of IEEE Spectrum magazines. The mean length of the conferences was 3.94 days, with a (sample) standard deviation of 1.28 days. Assume we are interested in estimating the mean conference length of all engineering conferences (population mean μ).

- What is the sampling distribution of \bar{X} = “sample mean length of the conferences” ? Why?

- Answer:** We **don't know** the **population standard deviation**, so we have to assume that the population random variable X = “length of conferences” is normally distributed with unknown population mean μ (population mean length of the conferences) and unknown standard deviation σ .

$$\bar{X} \sim N(\mu, \frac{s}{\sqrt{n}}) \sim N(\mu, \frac{1.28}{\sqrt{84}}) \sim N(\mu, 0.1397)$$

- What is the EBM for a 95% CI for μ ? $EBM = z_{\alpha/2} * \frac{s}{\sqrt{n}} = z_{0.025} * 0.1397 = 1.96 * 0.1397 = 0.274$
- What is the 95% CI for μ ? 3.94 ± 0.274
- What is the 99% CI for μ ? $3.94 \pm z_{0.005} * 0.1397 = 3.94 \pm 2.576 * 0.1397 = 3.94 \pm 0.36$

You have more confidence, but the confidence interval is wider (less precise)

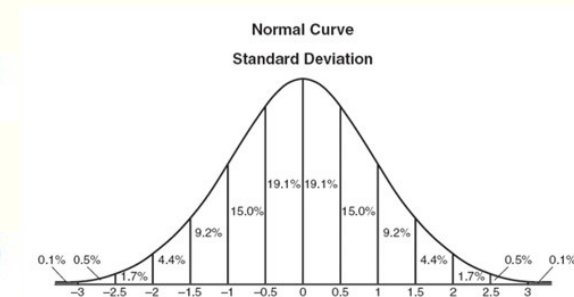
- What does the 99% CI for μ say?

It says that for 99% of random samples of the same size, the interval $\bar{X} \pm 2.576 * \frac{s}{\sqrt{84}}$ will contain the true population (unknown) mean μ .

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with $(n-1)$ degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size $<< 30$, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)

```
> qnorm(0.975,0,1)
[1] 1.959964
```

```
> qnorm(0.005,0,1)
[1] -2.575829
```

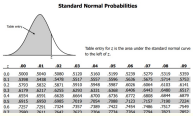


Book Problem - 8.110 - page 483

Practice with calculator



Practice with Tables



- Forbes magazine published data on the best small firms in 2012. These were firms that had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. The Table below shows the ages of the corporate CEOs for a random sample of these firms. Assume the underlying population is normal.

48	58	51	61	56	59	74	63	53	50
59	60	60	57	46	55	63	57	47	55
57	43	61	62	49	67	67	55	55	49

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with (n-1) degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{s}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size $\ll 30$, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)

- What is the sampling distribution of \bar{X} = “sample mean age of CEOs of best small firms of 2012”? Why?

This is Case 2 with the population normal, population standard deviation unknown and sample size = 30. You can use either the z-distribution i.e., $\bar{X} \sim \text{Normal}(\mu, s/\sqrt{n})$ or

$$\bar{X} = 56.57, s = 6.91 \text{ and } \frac{s}{\sqrt{n}} = \frac{6.83}{\sqrt{30}} = 1.247 ; \text{ So, } \bar{X} \sim \text{Normal}(\mu, 1.247)$$

- a. What is a 99% CI for the mean age of CEOs of the best small firms of 2012?

$$\bar{X} \pm z_{0.005} * \frac{s}{\sqrt{n}} = 56.57 \pm 2.576 * 1.247 = 56.57 \pm 3.212$$

$$\bar{X} \pm t_{0.005, 29} * \frac{s}{\sqrt{n}} = 56.57 \pm 2.756 * 1.247 = 56.86 \pm 3.437$$

```
> qnorm(0.005, 0, 1)
[1] -2.575829
```

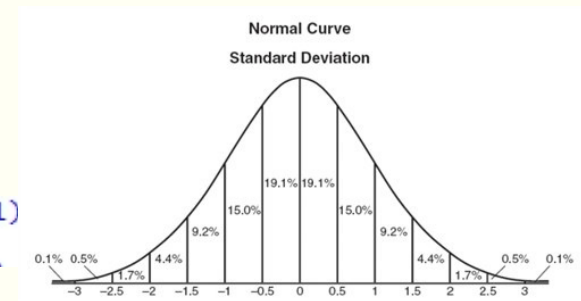
```
> print(qt(0.005, 29))
[1] -2.756386
```

- b. If we wanted an EBM that is half of what we have (i.e., doubly precise interval estimate) , with the same 99% confidence level, what should the sample size be?

Current EBM = 3.437 so we want EBM = 1.718.

That is, $t_{0.005, 29} * \frac{s}{\sqrt{n}} = 2.756 * \frac{6.83}{\sqrt{n}} = 1.718$.

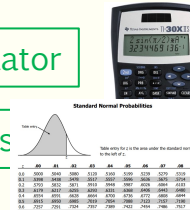
So, $\sqrt{n} = 2.756 * 6.83 / 1.718 = 10.96$ (approx. 11). So sample size n should be $11^2 = 121$



Book Problem - Modified Version of 8.101 - page 481

Practice with calculator

Practice with Tables



- The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. For a random selection of 40 House candidates $\bar{X} = \$568,873$. The population standard deviation for this data to the nearest hundred is $\sigma = \$909,200$.

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with (n-1) degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size < 30, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)

- What is the sampling distribution of \bar{X} = “sample mean of campaign contributions” ? Why?

- This is Case 3 with the population distribution unknown population standard deviation known. Sample size is 40.

So we use the Central Limit Theorem. $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ i.e., $\bar{X} \sim \text{Normal}(\mu, \frac{909200}{\sqrt{40}})$

- What is a 90% CI for the mean campaign contributions?

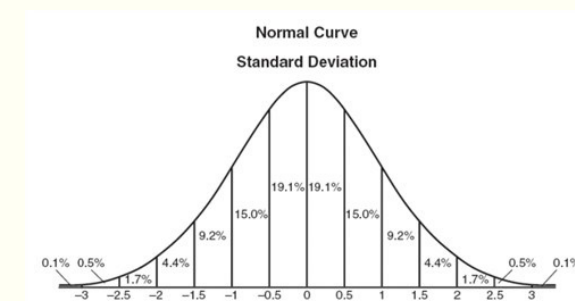
$$\text{EBM} = z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} = z_{0.05} * \frac{909200}{\sqrt{40}} = 1.645 * 143757 = 236480$$

$$\text{So } 90\% \text{ CI} = 568,873 \pm 236,480$$

```
> qnorm(0.05,0,1)
[1] -1.644854
```

- Interpret the Confidence Interval.

- It says that for 90% of random samples of the same size, the interval $\bar{X} \pm 236,480$ will contain the true population (unknown) mean μ .

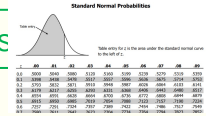


Problems

Practice with calculator



Practice with Tables



- Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

- $p' = 0.015$ is the sample proportion with $n = 451$. The sampling distribution of the sample proportion is: $p' \sim N(p, \sqrt{\frac{p(1-p)}{n}})$

- Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight-year period.

The error bound for a $(1 - \alpha)\%$ CI is $z_{\alpha/2} * \sqrt{\frac{p'(1-p')}{n}}$. We want 97% so $\alpha = 0.015$ and $\alpha/2 = 0.0075$.

$z_{\alpha/2} = z_{0.0075} = 2.43$ so that error bound = $2.43 * \sqrt{\frac{0.15 * 0.85}{451}} = 0.0357$.

So, 97% CI = 0.015 ± 0.0357

```
> qnorm(0.0075,0,1)
[1] -2.432379
```

- Interpret the above confidence interval

It says that for 97% of random samples of the same size, the interval $p' \pm 2.43 * \sqrt{\frac{p'(1-p')}{n}}$ will contain the true population (unknown) proportion p

- (i.e., population proportion of people over age 50 who ran and died in the same 8 year-period)

	Population σ known	Population σ unknown
Population Normal	$\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$. Use the standard normal z-distribution regardless of sample size. Case 1	<ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size ≥ 30. Use z-distribution. $\bar{X} \sim t(\mu, \frac{s}{\sqrt{n}})$ with $(n-1)$ degrees of freedom if sample size < 30. Use t-distribution. Case 2
Population Not Known Normal	CLT says: <ul style="list-style-type: none"> $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma}{\sqrt{n}})$ if sample size 30 or more. Use z-distribution <i>If sample size < 30, we really have to assume the population distribution. To avoid this, collect a larger sample. Case 3</i>	Obtain sampling distribution of \bar{X} using Bootstrapping. (we will see later)

