Probability and Statstics

Dr. Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

Textbook

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

Reference books

- **Probability Demystified**, Allan G. Bluman
- **Schaum's Outline of Probability and Statistics**
- **MATLAB Primer**, Seventh Edition
- **MATLAB Demystified** by McMahon, David

References

Readings for these lecture notes:

- **Schaum's Outline of Probability, Second Edition (Schaum's Outlines)** by by Seymour Lipschutz, Marc Lipson
- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- https://wordwatchtowers.wordpress.com/2009/12/21/underestimate-or-overestimate/
- Elementary Statistics, Tenth Edition, Mario F. Triola
- http://www.sjsu.edu/faculty/gerstman/

These notes contain material from the above resources.

Populations and Samples

- The totality of observations with which we are concerned, whether their number be finite or infinite, constitutes what we call a **population**.

- A **population** consists of the totality of the observations with which we are concerned.

- A **sample** is a subset of a population.

Bias

Any **sampling procedure** that produces inferences that consistently **overestimate** or consistently **underestimate** some characteristic of the population is said to be **biased**.

To eliminate any **possibility of bias** in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made **independently** and at **random**.

Overestimate vs. Underestimate

- **Overestimate** means ‘to form too high an estimate of’

- **Underestimate** means to estimate that something is smaller or less important than it actually is

Parameter vs. Statistic

- **Statistical inference** involves drawing conclusions about **characteristics of populations**.

- Among these characteristics are constants which are called **population parameters**. Two important parameters are the **population mean** and the **population variance**.

- Any function of the random variables constituting a **random sample** is called a **statistic**.

Sampling Distribution [1]

- The probability distribution of a statistic is called a **sampling distribution**.

- The field of **statistical inference** is basically concerned with **generalizations** and **predictions**.

- For example, we might claim, based on the opinions of several people interviewed on the street, that in a forthcoming election **60% of the eligible voters** in the city of Detroit favor a certain candidate. In this case, **we are dealing with a random sample of opinions from a very large finite population**.

Sampling Distribution [2]

- As a second illustration we might state that the **average cost** to build a residence in Charleston, South Carolina, is between **$330,000 and $335,000**, based on the **estimates of 3 contractors** selected at random from the 30 now building in this city. The population being sampled here is **again finite** but **very small**.

Sampling Distribution [3]

- Finally, let us consider a **soft-drink machine** designed to dispense, on average, **240 milliliters** per drink. A company official who computes the mean of **40** drinks obtains $\bar{x}$**= 236** milliliters and, on the basis of this value, decides that the machine is still dispensing drinks with an average content of μ **= 240** milliliters. The **40** drinks represent a sample from the **infinite population** of possible drinks that will be dispensed by this machine.

Inference about the Population from Sample Information [1]

- In each of the examples above, we computed a **statistic** from a **sample selected** from the **population**, and from this **statistic** we made various statements concerning the values of **population parameters** that **may or may not be true**.
- The company official made the decision that the soft-drink machine dispenses drinks with an average content of **240 milliliters**, even though the sample mean was **236 milliliters**, because he knows from sampling theory that, if **$\mu = 240$** milliliters, such a sample value could easily occur.

Inference about the Population from Sample Information [2]

- In fact, if he ran similar tests, say every hour, he would expect the values of the statistic $\bar{x}$ **to fluctuate** above and below $\mu = 240$ milliliters. Only when the value of $\bar{x}$ is substantially different from 240 milliliters will the company official initiate action to adjust the machine.

- Since a statistic is a **random variable** that depends only on the observed sample, it must have a **probability distribution**.

- The probability distribution of a **statistic** is called a **sampling distribution**.

Sampling Distribution of a Statistic

The **sampling distribution of a statistic** (such as a **sample proportion** or **sample mean**) is the distribution of all values of the statistic when **all possible samples** of the **same size *n*** are taken from the same population.

Sampling Distribution of the Mean

The **sampling distribution of the mean** is the distribution of **sample means**, with **all samples** having the **same sample size *n*** taken from the **same population**.

Sampling Distribution of the Proportion

- The **sampling distribution of the proportion** is the distribution of **sample proportions**, with all samples having the **same sample size *n*** taken from the **same population**.

The Central Limit Theorem

Central Limit Theorem: If $\bar{X}$ is the mean of a random **sample of size *n*** taken from a **population** with **mean μ** and **finite variance σ^2**, then the limiting form of

the distribution of $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$,

as $n \to \infty$, is the **standard normal distribution $n(z; 0, 1)$.**

The Central Limit Theorem

- The normal approximation for $\bar{X}$ will generally be good if ***n* ≥ 30**, provided the population distribution is not **terribly skewed**.
- If ***n* < 30**, the approximation is good only if the population is **not too different** from a **normal distribution**.
- As stated above, if the **population is known to be normal**, the **sampling distribution of** $\bar{X}$ will follow a **normal distribution** exactly, no matter how **small the size of the samples**.

The Central Limit Theorem

- The sample size ***n* = 30** is a guideline to use for the **Central Limit Theorem**.

- However, as the statement of the theorem implies, the presumption of normality on the distribution **of** $\overline{X}$ becomes more accurate **as *n* grows larger**.

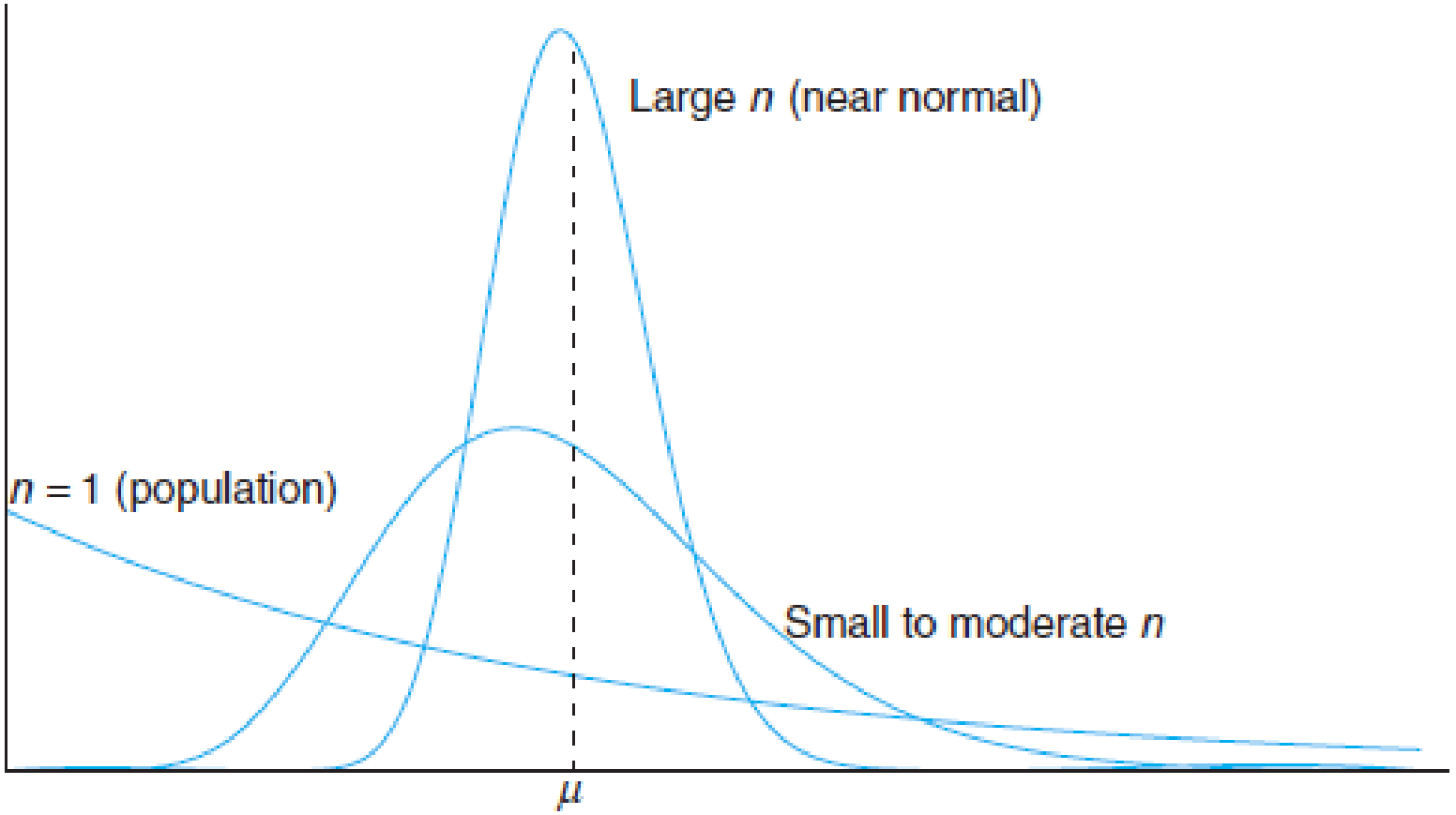


Illustration of the Central Limit Theorem (distribution of $\bar{X}$ for n = 1, moderate n, and large n).

The Central Limit Theorem

When selecting a **simple random sample** from a **population** with **mean** and **standard deviation**, it is essential to know these principles:

1. If ***n* > 30**, then the **sample means** have a distribution that can be **approximated by a normal distribution** with **mean μ** and **standard deviation σ /$\sqrt{n}$** (This guideline is commonly used, regardless of the distribution of the original population.)

2. If ***n* ≤ 30** and the **original population** has a **normal distribution**, then the **sample means** have a normal distribution with mean **μ** and standard deviation **σ /$\sqrt{n}$**

The Central Limit Theorem

3. If $n \leq 30$ but the **original population** does not have **a normal distribution**, then the methods of this section do not apply.

- Try to keep this big picture in mind: As we sample from a population, we want to know the behavior of the sample means.

- The ***central limit theorem*** tells us that if the **sample size is large enough**, the distribution of **sample means** can be **approximated by a *normal distribution***, even if the **original population is not** normally distributed.

The Central Limit Theorem and the Sampling Distribution of $\overline{x}$

Given:

1. The **random variable *x*** has a **distribution (which may or may not be normal)** with mean **μ** and standard deviation **σ**.

2. **Simple random samples** all of the **same size *n* are selected from the population**. (The samples are selected so that all possible samples of size *n* have the same chance of being selected.)

The Central Limit Theorem and the Sampling Distribution of $\overline{x}$

Conclusions:

1. The distribution of **sample means** $\overline{x}$ will, as the sample size increases, approach a ***normal distribution***.
2. The **mean of all sample means** is the population mean (That is, the normal distribution from Conclusion 1 has mean **μ**).
3. The **standard deviation** of all **sample means** is $\sigma/\sqrt{n}$ (That is, the normal distribution from Conclusion 1 has standard deviation $\sigma/\sqrt{n}$)

The Central Limit Theorem and the Sampling Distribution of $\overline{x}$

Practical Rules Commonly Used

If the original population is not itself normally distributed, here is a common guideline:

1. For samples of **size *n* greater than 30**, the distribution of the **sample means** can be **approximated reasonably well** by a **normal distribution**.(There are **exceptions**, such as **populations** with very **nonnormal** distributions requiring sample **sizes larger than 30**, but such exceptions **are relatively rare**.) **The approximation gets better as the sample size *n* becomes larger**.

The Central Limit Theorem and the Sampling Distribution of $\overline{x}$

Practical Rules Commonly Used

2. If the **original population** is itself **normally distributed**, then the **sample means** will be **normally distributed** for ***any* sample size *n*** (**not just the values of *n* larger than 30**).

Notation for Sampling Distribution of $\overline{x}$

The **central limit theorem** involves **two different distributions**: the distribution of the **original population** and the distribution of the s**ample means**.

- We use the symbols and to denote the mean **μ** and standard deviation **σ** of the original population, but we use the following **new notation** for the **mean** and **standard deviation** of the **distribution of sample means.**

$\mu_{\overline{X}} = \mu$ *and*

$\sigma_{\overline{X}} = \sigma / \sqrt{n}$

Example Water Taxi Safety

We noted that some passengers died when a water taxi sank in Baltimore's Inner Harbor. **Men are typically heaver than women and children, so when loading a water taxi, let's assume a worst-case scenario in which all passengers are men**. Based on data from the National Health and Nutrition Examination Survey, assume that weights of men are normally distributed with a **mean** of **172 lb** and a **standard deviation of 29** lb.

Example Water Taxi Safety cont.

a. Find the probability that if an individual man is randomly selected, his weight will be **greater than 175 lb**.

b. Find the probability that **20** randomly selected men will have a **mean that is greater than 175 lb** (so that their total weight exceeds the safe capacity of 3500 lb).

Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Area under the Normal Curve [2]

Table A.3 (continued) Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	0.5517	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	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

μ = 172 and **σ = 29**

a) $Z = \frac{x - \mu}{\sigma} = Z = \frac{175 - 172}{29} =$ **0.10**

P (X > 175) = P (Z > 0.10) = 1 - P (Z < 0.10) = 1 – 0. 5398

P (X > 175) = **0.4602 ans**

b) $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{175 - 172}{29/\sqrt{20}} = \frac{3}{6.4846} =$ **0.46**

P ($\bar{x}$ > 175) = P (z > 0.46) = 1- P ($z \leq 0.46$)

= 1 – 0.6772

P ($\bar{x}$ > 175) = **0.3228 ans**

Basics of inference[1]

- **Statistical inference** is the act of **generalizing** from a **sample to a population** with calculated degree of certainty. The two forms of statistical inference are **estimation** and **hypothesis testing**.

- A statistical **population** represents the set of all possible values for a variable. In practice, we do not study the entire population.

Basics of inference[2]

- Instead, we use data in a **sample** to shed light on the wider population.

- The term **parameter** is used to refer to a numerical **characteristic** of a **population**. Examples of parameters include the **population mean (μ)** and the population **standard deviation (σ)**.

Basics of inference[3]

- A numerical characteristic of the **sample** is a statistic.

- We introduce a particular type of **statistic** called an **estimate**. The **sample mean** $\bar{x}$ is the natural estimator of **population mean μ**. Sample standard **deviation s** is the natural estimator of population **standard deviation σ**.

- The parameter is a fixed **constant**. In contrast, the **estimator varies** from **sample to sample**.

Basics of inference[4]

	Parameter	Estimators
Source	Population	Sample
Value known?	No	Yes (calculate)
Notation	Greek (μ)	Roman ($\bar{x}$)
Vary from sample to sample	No	Yes
Error-prone	No	Yes

Sampling distribution of a mean (SDM)

- If we had the opportunity to take repeated samples from the same population, samples means ($\bar{x}$s) would vary from sample to sample and form a **sampling distribution means (SDM)**.

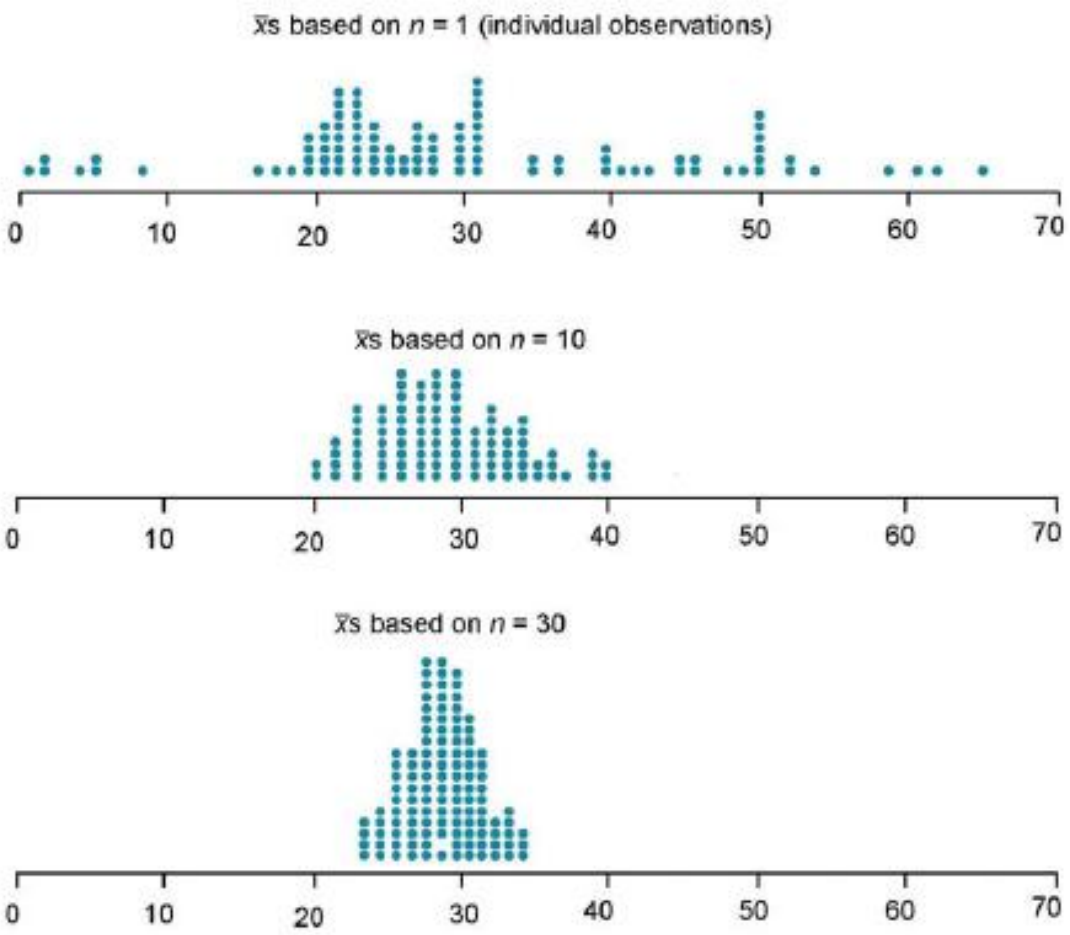
- Let's run a simulation experiment. Our simulation will be based on sampling a population of **N = 600** age values. The population mean **age μ = 29.5.** The population standard deviation **σ = 13.6**

Sampling distribution of a mean (SDM)

- Imagine taking repeated samples, each of ***n* = 10**. Do this **100 times**.

- In one experiment, it just so happened that the first $\bar{x}$ was **36.4**, the second $\bar{x}$ was **30.2**, and the third $\bar{x}$ was **24.6**

Sampling distribution of $\overline{x}$



Statistical Inference

- **Statistical inference** consists of those methods by which one makes **inferences or generalizations** about a **population**.

- The trend today is to distinguish between the **classical method** of estimating a **population parameter**, whereby inferences are based strictly on information obtained from a **random sample** selected from the **population**.

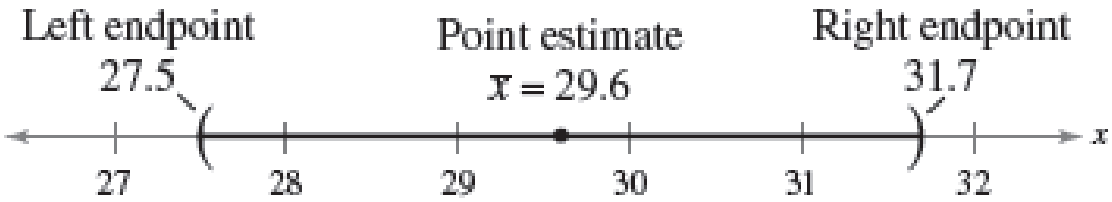
- **Statistical inference** may be divided into two major areas: **estimation** and **tests of hypotheses**.

Point Estimate

A **point estimate** is a single value estimate for a population parameter. The most unbiased point estimate of the population mean **μ** is the sample mean $\overline{x}$.

Interval Estimate

An **interval estimate** is an interval, or range of values, used to estimate a population parameter.



Interval Estimate

To form an **interval estimate**, use the **point estimate** as the center of the interval, and then add and subtract a **margin of error**.

Level of Confidence

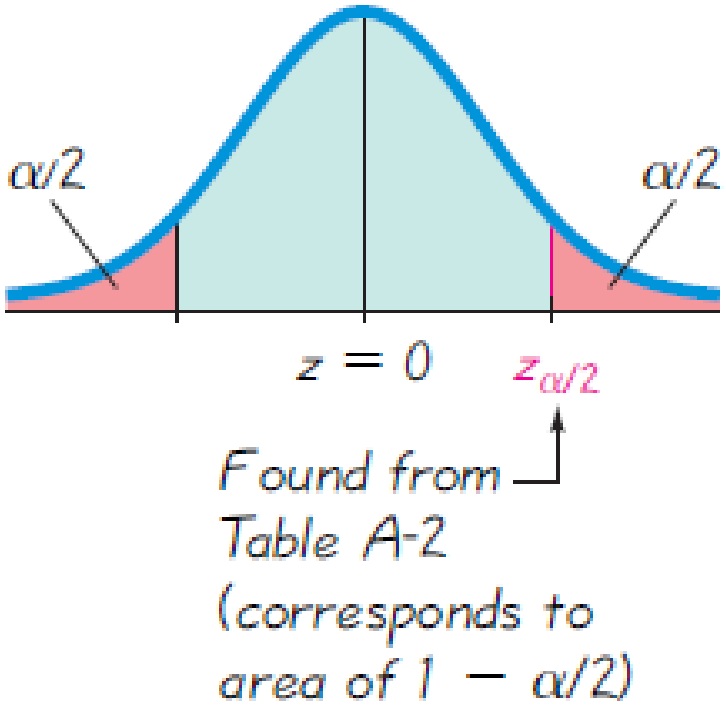
The **level of confidence *c* *or 1 - α*** is the probability that the **interval estimate** contains the population parameter, assuming that the estimation process is **repeated a large number of times**.

Critical Values [1]

- A **critical value** is the number on the borderline separating sample statistics that **are likely** to occur from those that are **unlikely to occur.**

- The number $z_{\alpha/2}$ is a **critical value** that is a ***z* score** with the property that it separates an area of **$\alpha/2$** in the right tail of the standard normal distribution.

Critical Values[2]



Critical Value $z_{\alpha/2}$ in the Standard Normal Distribution

Example Finding a Critical Value Find the critical value $z_{\alpha/2}$ corresponding to a 95% confidence level.

Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Area under the Normal Curve [2]

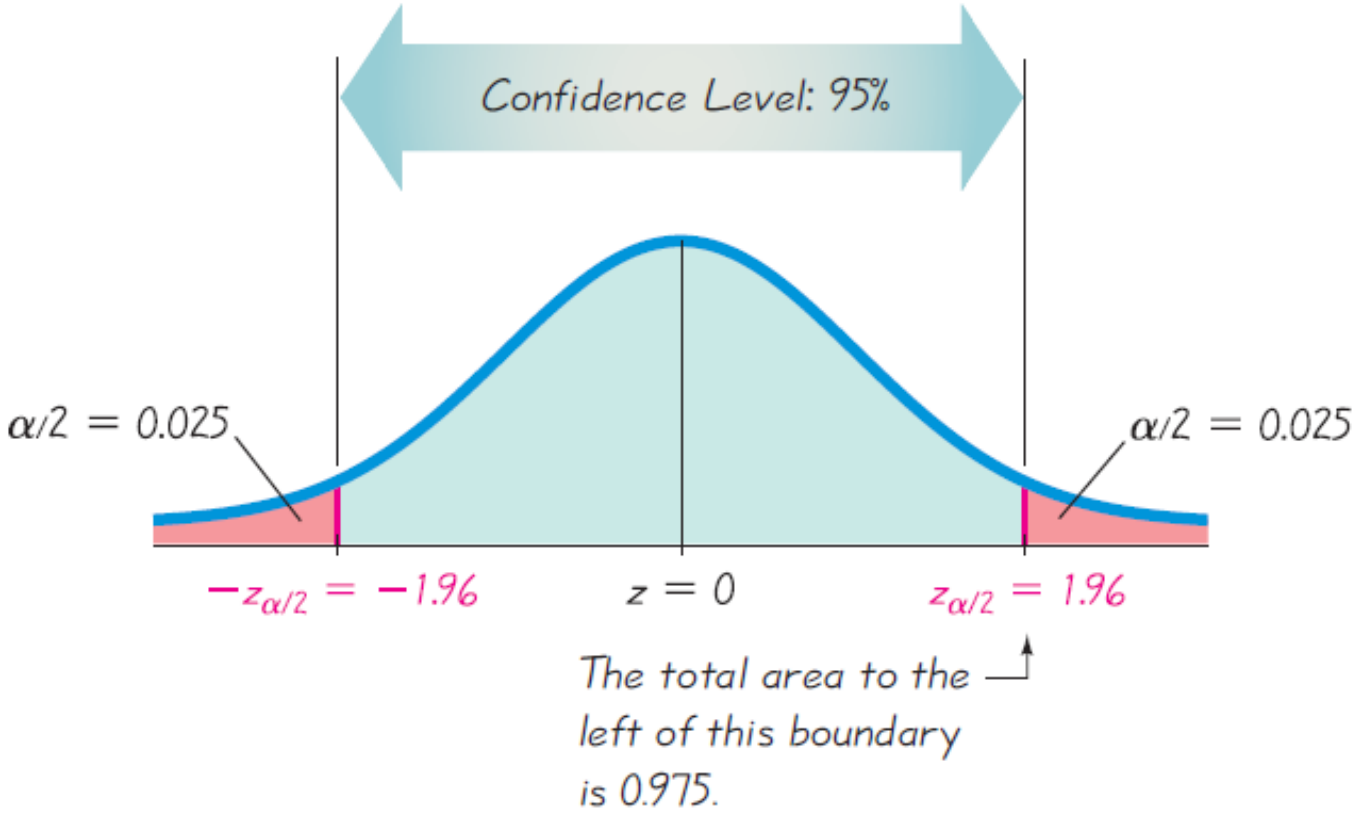
Table A.3 (continued) Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	0.5517	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	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Solution

When $\alpha = 0.05$

$z_{\alpha/2} = z_{0.0250} = 1.96$ $\quad \because 1 - \alpha/2 = 1 - 0.250 = 0.9750$



Confidence Level (1 – α)	α	Critical values, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

Margin of Error

Given a level of confidence $1 - \alpha$, the **margin of error** E (sometimes also called the **maximum error of estimate** or error tolerance) is the **greatest possible distance** between the **point estimate** and the **value of the parameter it is estimating**

$E = Z_{\alpha} \frac{\sigma}{\sqrt{n}}$ Margin of error for **μ** (σ known)

when these conditions are met.

1. The sample is **random**.

2. At least one of the following is true: The population is **normally distributed** or $n \geq 30$.

Estimation

A candidate for public office may wish to **estimate** the true **proportion** of voters favoring him by obtaining opinions from a **random sample of 100** eligible voters.

The fraction of voters in the sample favoring the candidate could be used as an estimate of the **true proportion in the population** of voters.

A knowledge of the **sampling distribution** of a **proportion** enables one to **establish the degree of accuracy** of such an **estimate**. This problem falls in the **area of estimation**.

Tests of Hypotheses

A floor wax is more scuff-resistant than brand *B* floor wax. He or she might hypothesize that **brand *A* is better than brand** *B* and, after proper testing, accept or reject this hypothesis.

In this example, we do not attempt to **estimate a parameter**, but instead we try to arrive at a correct decision about a **prestated hypothesis**.

Once again we are dependent on **sampling theory** and the use of data to provide us with some measure of accuracy for our decision.

Point Estimate [1]

A **point estimate** of some population parameter θ is a single value of a statistic $\hat{\theta}$.

For example, the value $\bar{x}$ of the statistic $\bar{X}$, computed from a sample of size n, is a point estimate of the population parameter μ. Similarly, $\hat{p} = x/n$ is **a point estimate** of the **true proportion p** for a binomial experiment.

An estimator is not expected to estimate the population parameter **without error**. We do not expect $\bar{X}$ to estimate **μ** exactly, but we certainly hope that **it is not far off**.

Point Estimate[2]

For a particular sample, it is possible to obtain a closer estimate of **μ** by using the sample median $\tilde{X}$ as an estimator. Consider, for instance, a sample consisting of the values **2, 5, and 11** from a population whose **mean is 4** but is supposedly unknown.

We would estimate μ to **be** $\bar{x} = 6$, using the sample mean as our estimate, or $\tilde{x} = 5$, using the sample median as our estimate. In this case, the estimator $\tilde{X}$ produces an estimate closer to the true parameter than does the estimator $\bar{X}$.

Point Estimate[3]

On the other hand, if our random sample contains the values **2, 6, and 7**, then $\tilde{x} = 6$ and $\bar{x} = 5$, so $\bar{x}$ is the better estimator.

Not knowing the true value of μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as our estimator.

Unbiased Estimator

A statistic $\widehat{\theta}$ is said to be an **unbiased estimator** of the parameter θ if $\mu_{\widehat{\theta}} = E(\widehat{\theta}) = \theta$

Unbiased Estimator

Example 1: $E(\bar{x}) = \mu$, so $\bar{x}$ is an **unbiased estimator** of μ

Example 2: $E(s^2) = \sigma^2$, so s^2 is an **unbiased estimator** of σ^2

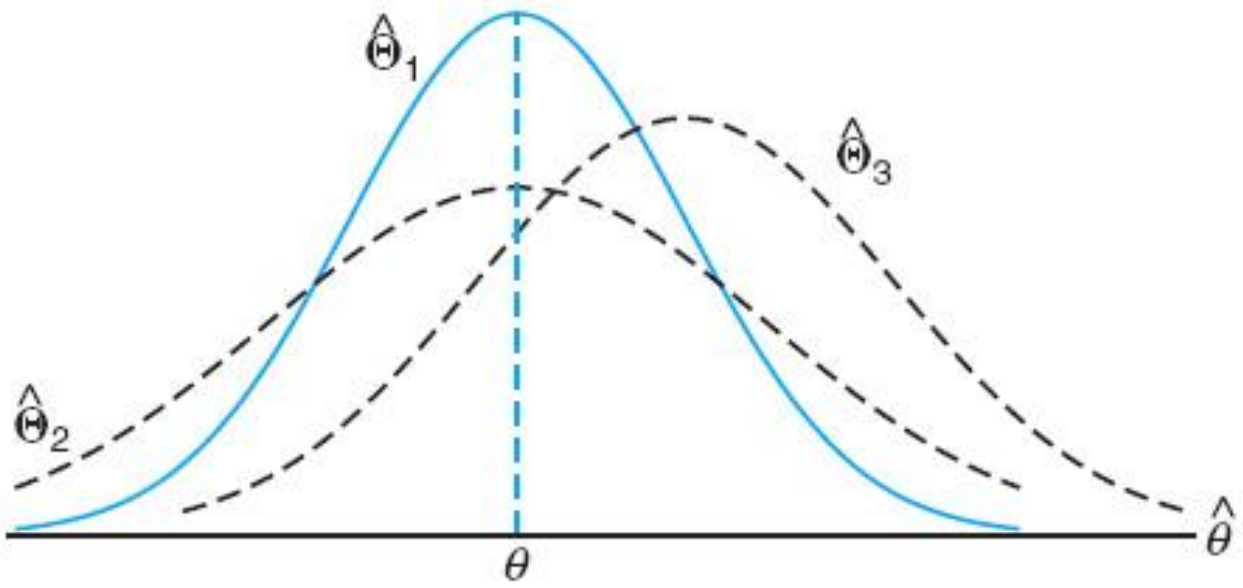
Variance of a Point Estimator [1]

- If $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are two unbiased estimators of the same population parameter θ, we want to choose the **estimator whose sampling distribution** has the **smaller variance**.

- Hence, if $\sigma^2_{\widehat{\theta_1}} < \sigma^2_{\widehat{\theta_2}}$, we say that $\widehat{\theta_1}$ is a **more efficient estimator** of θ than $\widehat{\theta_2}$.

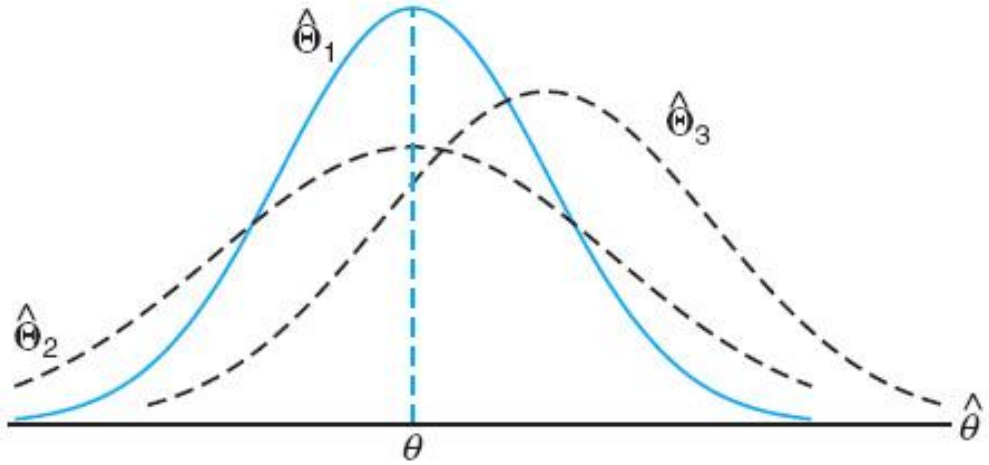
Variance of a Point Estimator [2]

- If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the **most efficient estimator** of **θ**.



Variance of a Point Estimator [3]

The figure illustrates the sampling distributions of three different estimators, $\widehat{\boldsymbol{\theta}_1}$, $\widehat{\boldsymbol{\theta}_2}$, and $\widehat{\boldsymbol{\theta}_3}$, all estimating $\boldsymbol{\theta}$. It is clear that only $\widehat{\boldsymbol{\theta}_1}$ **and** $\widehat{\boldsymbol{\theta}_2}$ are **unbiased**, since their distributions are centered at $\boldsymbol{\theta}$. The estimator $\widehat{\boldsymbol{\theta}_1}$has a smaller variance than $\widehat{\boldsymbol{\theta}_2}$ and is therefore more efficient. Hence, our choice for an estimator of $\boldsymbol{\theta}$, among the three considered, would be $\widehat{\boldsymbol{\theta}_1}$.



Variance of a Point Estimator [2]

For normal populations, one can show that both $\overline{X}$ **and** $\tilde{X}$ are **unbiased estimators** of the population mean μ, but the variance of $\overline{X}$ is smaller than the variance of $\tilde{X}$.

Thus, both estimates $\overline{x}$ and $\tilde{x}$ will, on average, equal the population mean μ, but $\overline{x}$ is likely to be closer to μ for a given sample, and thus $\overline{X}$ is **more efficient** than $\tilde{X}$