END 311E STATISTICS

Sampling and Sampling Distributions

By Çiğdem Kadaifçi Yanmaz Ph.D.

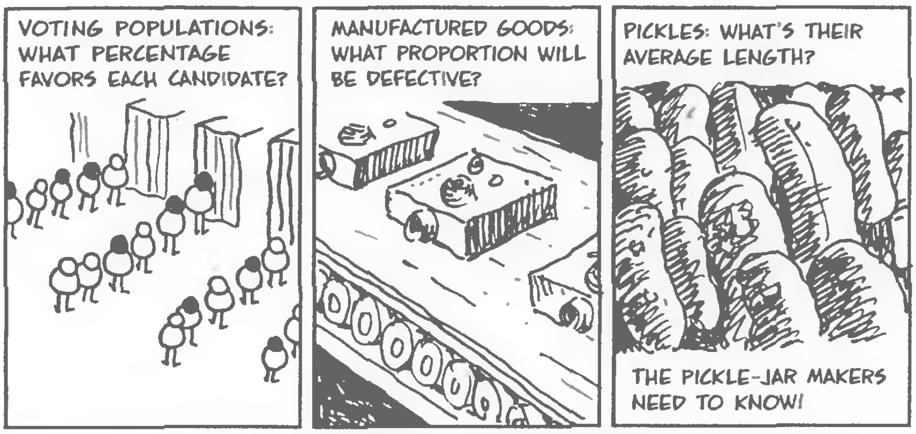
1

Population census

Are we able to observe all members of the population?

Motor vehicle registration

2



Sampling

- **Sampling:** The **process** or **action** of taking samples representing the population
- **Sample:** a finite **part of a population** whose properties are studied to gain information about the whole
- **Sample Statistics**, any value of observed data, especially one used to estimate the corresponding parameter of the population:

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

Why Sample?

- Selecting a sample is **less time-consuming** than selecting every item in the population.
- Selecting a sample is **less costly** than selecting every item in the population.
- An analysis of a sample is **less cumbersome** and **more practical** than an analysis of the entire population.

Sampling Distributions

- A sampling distribution is a distribution of all of the possible values of a sample statistic for a given size sample selected from a population.
- For example, suppose you sample 50 students from your college regarding their mean GPA. If you obtained many different samples of 50, you will compute a different mean for each sample. We are interested in the distribution of all potential mean GPA we might calculate for any given sample of 50 students.

Developing a Sampling Distribution

- **Assume there is a population**
- Population size $N = 4$
- Random variable, X, is age of individuals
- Values of X: 18, 20, 22, 24 (years)

16.11.2020 7

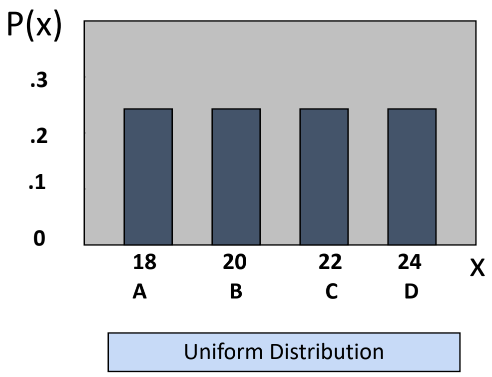
7

Developing a Sampling Distribution

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N} = \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



16.11.2020 8

8

Developing a Sampling Distribution

Now consider all possible samples of size n=2

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

16.11.2020 9

9

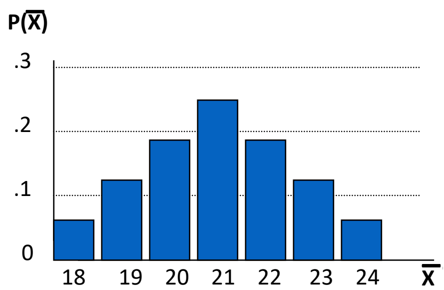
Developing a Sampling Distribution

Sampling Distribution of All Sample Means

16 Sample Means

	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



(no longer uniform)

16.11.2020

10

Developing a Sampling Distribution

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{X}} = \frac{18 + 19 + 19 + \cdots + 24}{16} = 21$$

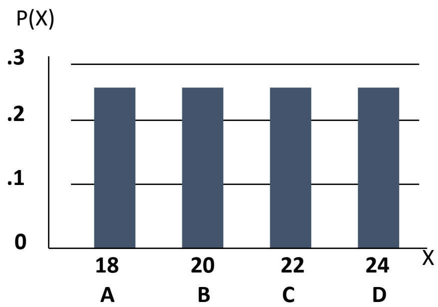
$$\sigma_{\bar{X}} = \sqrt{\frac{(18-21)^2 + (19-21)^2 + \cdots + (24-21)^2}{16}} = 1.58$$

Note: Here we divide by 16 because there are 16 different samples of size 2.

Comparing the Population Distribution to the Sample Means Distribution

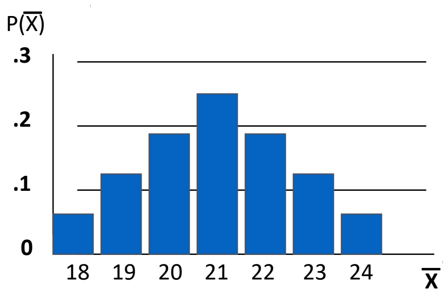
Population
N = 4

$\mu = 21 \quad \sigma = 2.236$



Sample Means Distribution
n = 2

$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$



Population mean = The mean of all possible samples

Sample mean is an UNBIASED ESTIMATOR for the population mean!!

Suppose we have random samples $X_1, X_2, \dots, X_n$

$$E\left[\sum_{i=1}^{n} X_i\right] = \underbrace{E[X_1]}_{\mu} + \underbrace{E[X_2]}_{\mu} + \cdots + \underbrace{E[X_n]}_{\mu} = n\mu$$

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n} n\mu = \mu$$

16.11.2020 13

13

Sample Mean Sampling Distribution: Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the **Standard Error of the Mean:**

  (This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases

16.11.2020 14

14

Sample variance is a BIASED ESTIMATOR for the population variance!!

Suppose we have random samples $X_1, X_2, \dots, X_n$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \underbrace{\mathrm{Var}(X_1)}_{\sigma} + \underbrace{\mathrm{Var}(X_2)}_{\sigma} + \cdots + \underbrace{\mathrm{Var}(X_n)}_{\sigma} = n\sigma^2$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

16.11.2020 15

15

Sample Mean Sampling Distribution: If the Population is Normal

- If a population is **normal** with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is **also normally distributed** with

$$\mu_{\bar{X}} = \mu$$ and $$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

16.11.2020 16

16

Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of $\bar{X}$:

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:
$\bar{X}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size

17

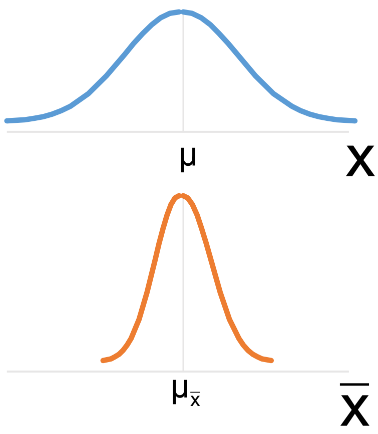
Sampling Distribution Properties

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

(i.e. $\bar{X}$ is unbiased)

Normal Population Distribution

Normal Sampling Distribution (has the same mean)



18

Example

- A cereal firm fills thousands of boxes of cereals during a day.
- To be consistent with the package labeling, boxes should contain 368 grams of cereal.
- Cereal weight varies from box to box.
- Given that the standard deviation of the cereal-filling process is 15 grams,
- What will the standard error be for a sample contains 25 boxes?

16.11.2020 19

19

Example

$$\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$$

The standard error is 3.

The variation in the sample means for samples of n = 25 is much less than the variation in the individual boxes of cereal

16.11.2020 20

20

Example

- What is the probability that the mean of a sample (n=25) being less than 365 gr?

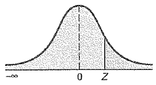
$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{365-368}{3} = -1$$

$$P\{Z < -1\} = 0.1587$$

TABLE E.2
The Cumulative Standardized Normal Distribution (*continued*)

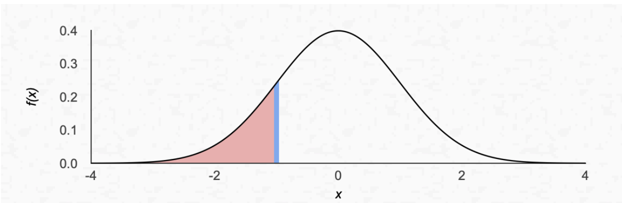
Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



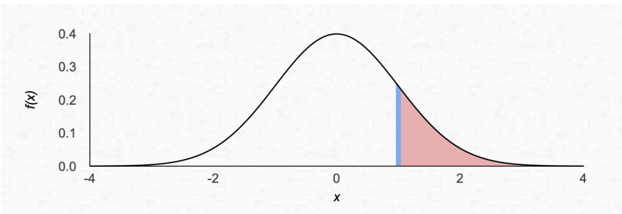
Cumulative Probabilities										
Z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	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	0.7518	0.7549
0.7	0.7580	0.7612	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

Example

$$P\{Z < -1\}$$

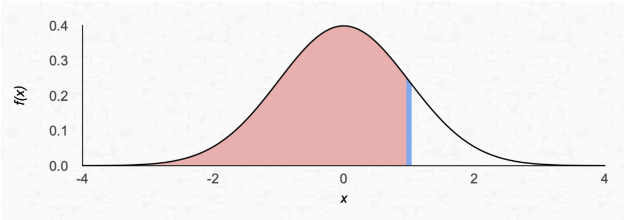


$$P\{Z > 1\}$$

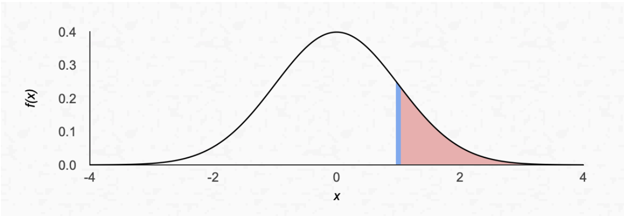


Example

$P\{Z < 1\}$



$1 - P\{Z < 1\}$



$P\{Z < -1\} = 1 - P\{Z < 1\}$

Example

- Find the probability that the mean being less than 365 gr using the population parameters.

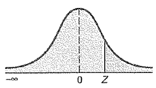
$$Z = \frac{365-368}{\mathbf{15}} = -0.2$$

$$\mathrm{P}\{Z < -0.2\} = 0.4207$$

many more individual boxes than sample means are below 365 grams

the chance that the sample mean of 25 boxes is far away from the population mean is less than the chance that a single box is far away!

TABLE E.2
The Cumulative Standardized Normal Distribution (*continued*)
Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



	Cumulative Probabilities									
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	0.5517	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	0.7518	0.7549
0.7	0.7580	0.7612	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

Example

taking a larger sample results in
less variability in the sample means from
sample to sample!

- What is the probability that the mean of a sample (n=100) being less than 365

$$\sigma_{\bar{X}} = \frac{15}{\sqrt{100}} = 1.5$$

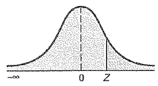
$$Z = \frac{365-368}{1.5} = -2$$

$$\mathrm{P}\{Z < -2\} = 0.0228$$

TABLE E.2

The Cumulative Standardized Normal Distribution (*continued*)

Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z

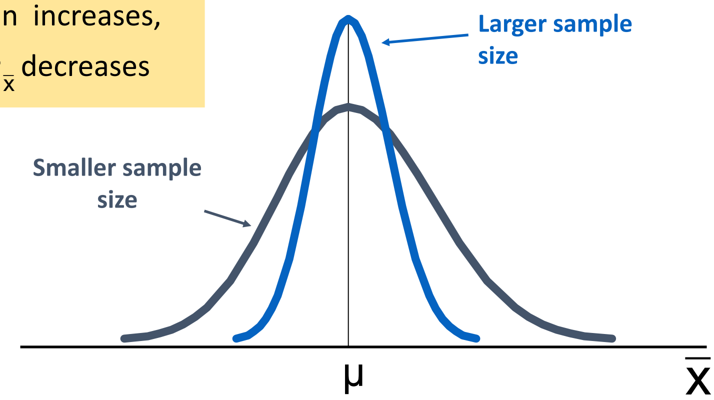


	Cumulative Probabilities									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	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	0.7518	0.7549
0.7	0.7580	0.7612	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

25

Sampling Distribution Properties

As n increases,
$\sigma_{\bar{x}}$ decreases



26

Determining An Interval Including A Fixed Proportion of the Sample Means

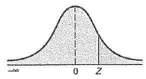
Find a symmetrically distributed interval around µ that will include 95% of the sample means when µ = 368, σ = 15, and n = 25.

- Since the interval contains 95% of the sample means 5% of the sample means will be outside the interval
- Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit.
- From the standardized normal table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96.

TABLE E.2

The Cumulative Standardized Normal Distribution (*continued*)

Entry represents area under the cumulative standardized normal distribution from $-\infty$ to Z



Cumulative Probabilities

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	0.5517	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	0.7518	0.7549
0.7	0.7580	0.7612	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

Determining An Interval Including A Fixed Proportion of the Sample Means

- Calculating the lower limit of the interval

$$\bar{X}_L = \mu + Z\frac{\sigma}{\sqrt{n}} = 368 + (-1.96)\frac{15}{\sqrt{25}} = 362.12$$

- Calculating the upper limit of the interval

$$\bar{X}_U = \mu + Z\frac{\sigma}{\sqrt{n}} = 368 + (1.96)\frac{15}{\sqrt{25}} = 373.88$$

- 95% of all sample means of sample size 25 are between 362.12 and 373.88

Sample Mean Sampling Distribution: If the Population is not Normal

- We can apply the **Central Limit Theorem**:
  - Even if the population is not normal,
  - ...sample means from the population will be approximately normal as long as the sample size is large enough.

Properties of the sampling distribution:

$\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

16.11.2020 29

29

$$E[S^2] = \sigma^2$$

$$E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{k}\right] = \sigma^2 \Rightarrow k = ?$$

$$\frac{1}{k}E[\sum_{i=1}^{n}\underbrace{(X_i - \bar{X})}_{(X_i - \mu) - (\bar{X} - \mu)}^2] = \sigma^2$$

$$(X_i - \mu) - (\bar{X} - \mu) \longrightarrow (X_i - \bar{X})$$

$$\frac{1}{k}E[\sum_{i=1}^{n}\left((X_i - \mu) - (\bar{X} - \mu)\right)^2] = \sigma^2$$

$$\frac{1}{k}E[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2] = \sigma^2$$

$$\frac{1}{k}\sum_{i=1}^{n}(E[(X_i - \mu)^2] - 2E[(X_i - \mu)(\bar{X} - \mu)] + E[(\bar{X} - \mu)^2]) = \sigma^2$$

30

$$\frac{1}{k}\sum_{i=1}^{n}(E[(X_i-\mu)^2]-2E[(X_i-\mu)(\bar{X}-\mu)]+E[(\bar{X}-\mu)^2])=\sigma^2$$

σ^2

The square of the standard error: $\frac{\sigma^2}{n}$

constant

$$\frac{1}{k}(n\sigma^2+n\frac{\sigma^2}{n}-2\frac{\boldsymbol{n}}{\boldsymbol{n}}E[\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu)])=\sigma^2$$

$$\frac{1}{k}(n\sigma^2+\sigma^2-2\frac{\boldsymbol{n}}{\boldsymbol{n}}E[(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)])=\sigma^2$$

$$\frac{1}{k}(n\sigma^2+\sigma^2-2\boldsymbol{n}E[(\bar{X}-\mu)\sum_{i=1}^{n}\frac{(X_i-\mu)}{\boldsymbol{n}}])=\sigma^2$$

31

$$\frac{1}{k}(n\sigma^2+\sigma^2-2\boldsymbol{n}E[(\bar{X}-\mu)\sum_{i=1}^{n}\frac{(X_i-\mu)}{\boldsymbol{n}}])=\sigma^2$$

$$\frac{1}{k}(n\sigma^2+\sigma^2-2\boldsymbol{n}\underbrace{E[(\bar{X}-\mu)^2]}_{\frac{\sigma^2}{n}})=\sigma^2$$

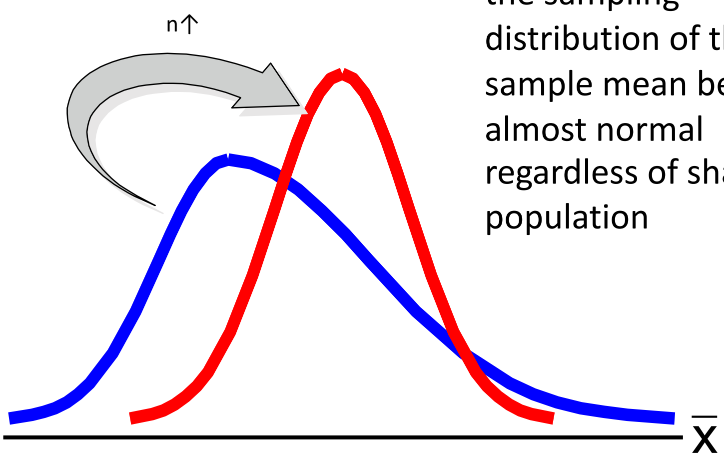
$$\frac{1}{k}(n\sigma^2+\sigma^2-2n\frac{\sigma^2}{n})=\sigma^2\Longrightarrow\frac{1}{k}(n\sigma^2+\sigma^2-2\sigma^2)=\sigma^2$$

$$\frac{1}{k}(n\sigma^2-\sigma^2)=\sigma^2\Longrightarrow\frac{1}{k}(n-1)\sigma^2=\sigma^2\Longrightarrow \boxed{k=(n-1)} \Longrightarrow \boxed{S=\sqrt{\frac{\sum_{i=1}^{n}(X_i-X)^2}{n-1}}}$$

32

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution of the sample mean becomes almost normal regardless of shape of population

33

Sample Mean Sampling Distribution: If the Population is not Normal

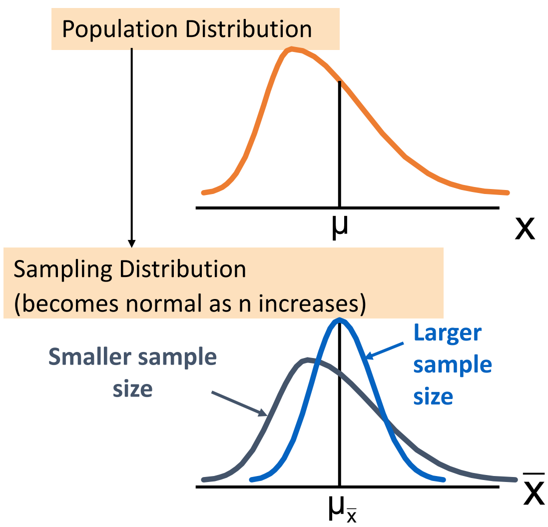
Sampling distribution properties:

Central Tendency

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

Variation

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



34

Example

- Suppose a population has mean μ = 8 and standard deviation σ = 3. Suppose a random sample of size n = 36 is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

Example

Solution:

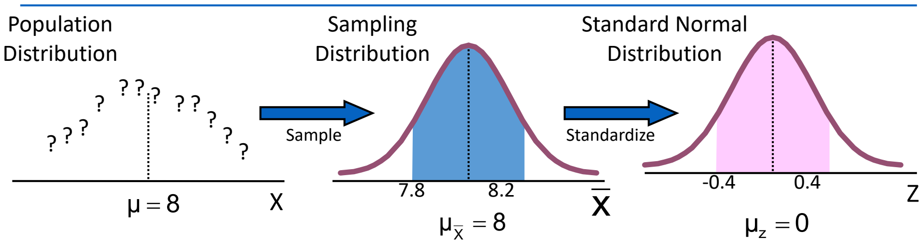
- Even if the population is not normally distributed, the central limit theorem can be used (n is relatively large)
- … so the sampling distribution of $\bar{X}$ is approximately normal
- … with $\mu_{\bar{x}} = 8$
- …and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

Example

Solution (continued):

$$P(7.8 < \bar{X} < 8.2) = P\left(\frac{7.8 - 8}{3/\sqrt{36}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}} \right)$$

$$= P(-0.4 < Z < 0.4) = 0.6554 - 0.3446 = 0.3108$$



37

Population Proportions

π = the proportion of the population having some characteristic

- Sample proportion (p) provides an estimate of π:

$$p = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

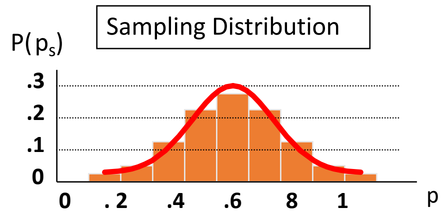
- $0 \leq p \leq 1$
- p is approximately distributed as a normal distribution when n is large

(assuming sampling with replacement from a finite population or without replacement from an infinite population)

38

Sampling Distribution of p

- Approximated by a normal distribution if:
  - $n\pi \geq 5$ and $n(1-\pi) \geq 5$



where

$\mu_p = \pi$ and $\sigma_p = \sqrt{\dfrac{\pi(1-\pi)}{n}}$

(where π = population proportion)

Standardize p to a Z value with the formula:

$$Z = \frac{p-\pi}{\sigma_p} = \frac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}$$

Suppose q is the probability of success and (1-q) is the probability of failure where $q = \frac{X}{n}$.

$E[X] = nq$
$\text{Var}[X] = nq(1-q)$

$$E[p] = E\left[\frac{X}{n}\right] = \frac{nq}{n} = q$$

$$\text{Var}[p] = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}nq(1-q) = \frac{q(1-q)}{n}$$

Example

- If the true proportion of voters who support Proposition A is $\pi = 0.4$, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?

i.e.: **if $\pi = 0.4$ and n = 200, what is**

$P(0.40 \leq p \leq 0.45)$?

41

Example

if $\pi = 0.4$ and $n = 200$, what is

$P(0.40 \leq p \leq 0.45)$?

Find σ_p:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$

Convert to standardized normal:

$$P(0.40 \leq p \leq 0.45) = P\left(\frac{0.40 - 0.40}{0.03464} \leq Z \leq \frac{0.45 - 0.40}{0.03464}\right)$$

$$= P(0 \leq Z \leq 1.44)$$

42

Example

if $\pi = 0.4$ and $n = 200$, what is
$P(0.40 \leq p \leq 0.45)$?

Utilize the cumulative normal table:

$P(0 \leq Z \leq 1.44) = 0.9251 - 0.5000 = 0.4251$