Chapter 8: Fundamental Sampling Distributions and Data Descriptions:

8.1 Random Sampling:

Definition 8.1:

A population consists of the totality of the observations with which we are concerned. (Population=Probability Distribution)

Definition 8.2:

A sample is a subset of a population.

Note:

- Each observation in a population is a value of a random variable X having some probability distribution f(x).
- To eliminate bias in the sampling procedure, we select a random sample in the sense that the observations are made independently and at random.
- The random sample of size *n* is:

$$X_1, X_2, \ldots, X_n$$

It consists of *n* observations selected independently and randomly from the population.

8.2 Some Important Statistics:

Definition 8.4:

Any function of the random sample $X_1, X_2, \ldots, X_n$ is called a statistic.

Central Tendency in the Sample:

Definition 8.5:

If $X_1, X_2, \ldots, X_n$ represents a random sample of size *n*, then the sample mean is defined to be the statistic:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(unit)}$$

Note:

- $\overline{X}$ is a statistic because it is a function of the random sample $X_1, X_2, \ldots, X_n$.
- $\overline{X}$ has same unit of $X_1, X_2, \ldots, X_n$.
- $\overline{X}$ measures the central tendency in the sample (location).

Variability in the Sample:

Definition 8.9:

If $X_1, X_2, \ldots, X_n$ represents a random sample of size n, then the sample variance is defined to be the statistic:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \text{ (unit)}^2$$

Theorem 8.1: (Computational Formulas for S^2)

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1} = \frac{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}{n(n-1)}$$

Note:

- S^2 is a statistic because it is a function of the random sample $X_1, X_2, \ldots, X_n$.
- S^2 measures the variability in the sample.

Definition 8.10:

The sample standard deviation is defined to be the statistic:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \text{ (unit)}$$

Example 8.1: Reading Assignment

Example 8.8: Reading Assignment

Example 8.9: Reading Assignment

8.4 Sampling distribution:

Definition 8.13:

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

- Example: If $X_1, X_2, \ldots, X_n$ represents a random sample of size n, then the probability distribution of $\overline{X}$ is called the sampling distribution of the sample mean $\overline{X}$.

8.5 Sampling Distributions of Means:

Result:

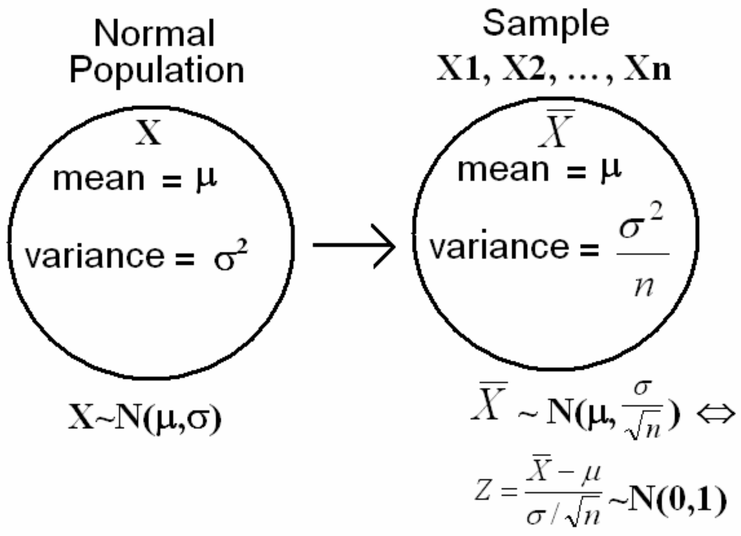
If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu,\sigma)$, then the sample mean $\overline{X}$ has a normal distribution with mean

$$E(\overline{X}) = \mu_{\overline{X}} = \mu$$

and variance

$$Var(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

- If X_1, X_2, …, X_n is a random sample of size n from N(μ,σ), then $\overline{X}$ ~N($\mu_{\overline{X}}, \sigma_{\overline{X}}$) or $\overline{X}$ ~N(μ,$\frac{\sigma}{\sqrt{n}}$).
- $\overline{X}$ ~ N(μ,$\frac{\sigma}{\sqrt{n}}$) ⇔ $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$~N(0,1)



Theorem 8.2: (Central Limit Theorem)
If X_1, X_2, …, X_n is a random sample of size n from any distribution (population) with mean μ and finite variance σ^2, then, if the sample size n is large, the random variable

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

is approximately standard normal random variable, i.e., $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$~N(0,1) approximately.

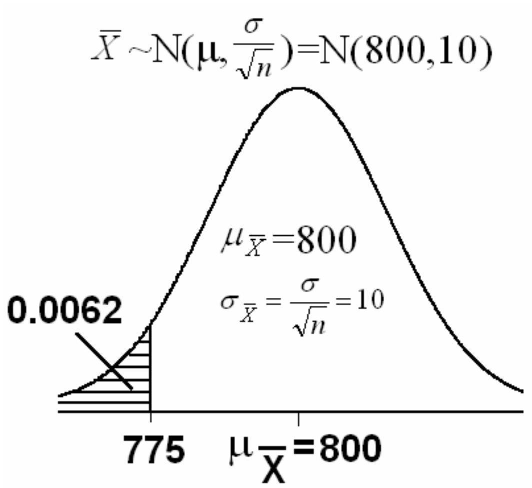
- $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$~N(0,1) ⇔ $\overline{X}$ ~ N(μ,$\frac{\sigma}{\sqrt{n}}$)
- We consider n large when $n \geq 30$.

- For large sample size n, $\overline{X}$ has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, i.e., $\overline{X}$ ~N(μ,$\frac{\sigma}{\sqrt{n}}$) approximately.
- The sampling distribution of $\overline{X}$ is used for inferences about the population mean μ.

Example 8.13:

An electric firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Solution:



X= the length of life

μ=800 , σ=40

X~N(800, 40)

n=16

$\mu_{\overline{X}} = \mu$=800

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$$

$\overline{X}$ ~ N(μ,$\frac{\sigma}{\sqrt{n}}$)=N(800,10)

$$\Leftrightarrow \quad Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = Z = \frac{\overline{X} - 800}{10} \sim \text{N}(0,1)$$

$$P(\overline{X} < 775) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{775 - \mu}{\sigma/\sqrt{n}}\right)$$

$$= P\left(\frac{\overline{X} - 800}{10} < \frac{775 - 800}{10}\right)$$

$$= P\left(Z < \frac{775 - 800}{10}\right)$$

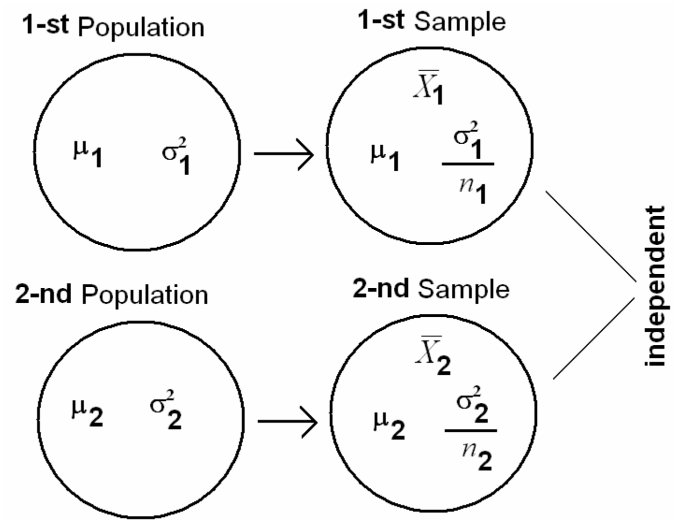
$$= P(Z < -2.50)$$

= 0.0062

Sampling Distribution of the Difference between Two Means:

Suppose that we have two populations:

- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about $\mu_1-\mu_2$.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\overline{X}_1$ be the sample mean of the 1-st sample.
- Let $\overline{X}_2$ be the sample mean of the 2-nd sample.
- The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is used to make inferences about $\mu_1-\mu_2$.



Theorem 8.3:

If n_1 and n_2 are large, then the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal with mean

$$E(\overline{X}_1 - \overline{X}_2) = \mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2$$

and variance

$$Var(\overline{X}_1 - \overline{X}_2) = \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

that is:

$$\bar{X}_1 - \bar{X}_2 \sim \text{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$\Leftrightarrow$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \text{N}(0,1)$$

Note:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1 - \bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \neq \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}} = \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$$

Example 8.15: Reading Assignment

Example 8.16:
The television picture tubes of manufacturer *A* have a mean lifetime of 6.5 years and standard deviation of 0.9 year, while those of manufacturer *B* have a mean lifetime of 6 years and standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer *A* will have a mean lifetime that is at least 1 year more than the mean lifetime of a random sample of 49 tubes from manufacturer *B*?

Solution:

Population A	Population B
μ_1=6.5	μ_2=6.0
σ_1=0.9	σ_2=0.8
n_1=36 (n_1>30)	n_2=49 (n_2>30)

- We need to find the probability that the mean lifetime of manufacturer *A* is at least 1 year more than the mean lifetime of manufacturer *B* which is P($\bar{X}_1 \geq \bar{X}_2 + 1$).
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

$$\bar{X}_1 - \bar{X}_2 \sim \text{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

- $E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$=6.5– 6.0=0.5
- $Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(0.9)^2}{36} + \frac{(0.8)^2}{49}$=0.03556

- $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}} = \sqrt{0.03556}$ = 0.189
- $\bar{X}_1 - \bar{X}_2$~N(0.5,0.189)
- Recall $Z = \dfrac{(\bar{X}_1 - \bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$~N(0,1)

P($\bar{X}_1 \geq \bar{X}_2 + 1$) = P($\bar{X}_1 - \bar{X}_2 \geq 1$)

$$= P\left(\frac{(\bar{X}_1 - \bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \geq \frac{1-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\right)$$

$$= P\left(Z \geq \frac{1-0.5}{0.189}\right)$$

= P(Z≥2.65)
= 1 – P(Z<2.65)
= 1 – 0.9960
= 0.0040

8.7 t-Distribution:

- Recall that, if $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e. N(μ,σ), then

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

- We can apply this result only when σ^2 is known!
- If σ^2 is unknown, we replace the population variance σ^2 with the sample variance $S^2 = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ to have the following statistic

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

Result:

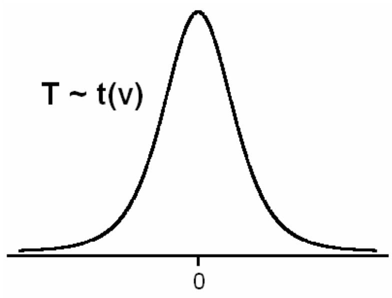
If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e. N(μ,σ), then the statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

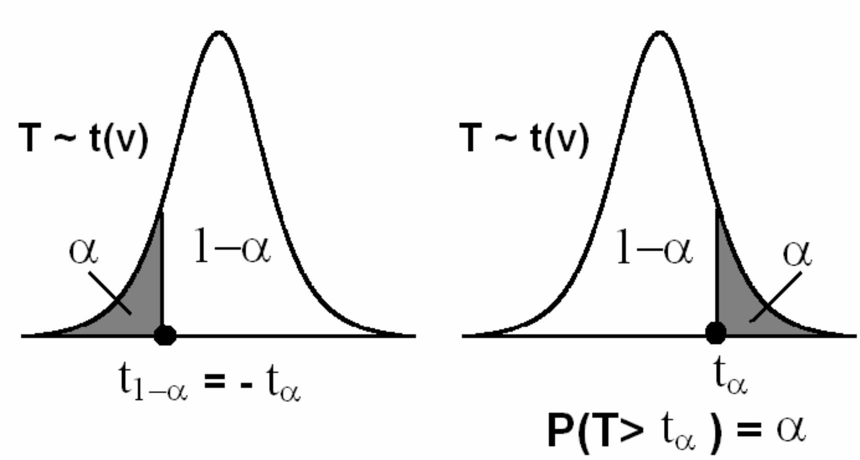
has a t-distribution with $\nu=n-1$degrees of freedom (df), and we write T~ t(ν) or T~ t(n–1).

Note:

- t-distribution is a continuous distribution.
- The shape of t-distribution is similar to the shape of the standard normal distribution.



Notation:



- t_α = The t-value above which we find an area equal to α, that is $P(T > t_\alpha) = \alpha$
- Since the curve of the pdf of T~ t(ν) is symmetric about 0, we have

$$t_{1-\alpha} = -t_\alpha$$

- Values of t_α are tabulated in Table A-4 (p.683).

Example:

Find the t-value with ν=14 (df) that leaves an area of:

(a) 0.95 to the left.

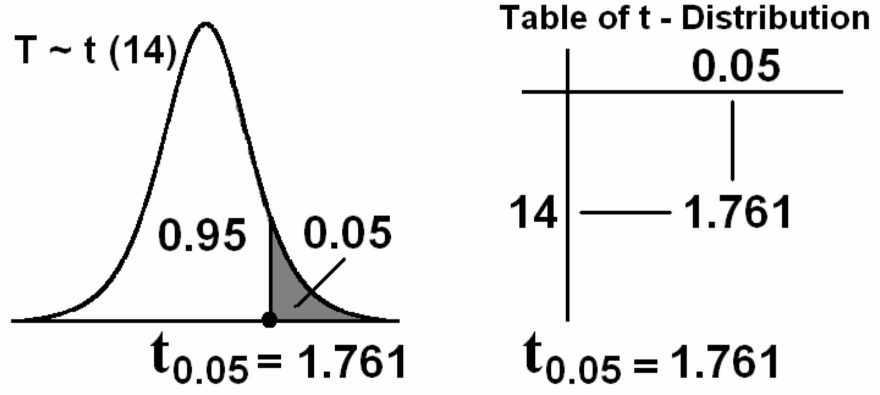
(b) 0.95 to the right.

Solution:

ν = 14 (df); T~ t(14)

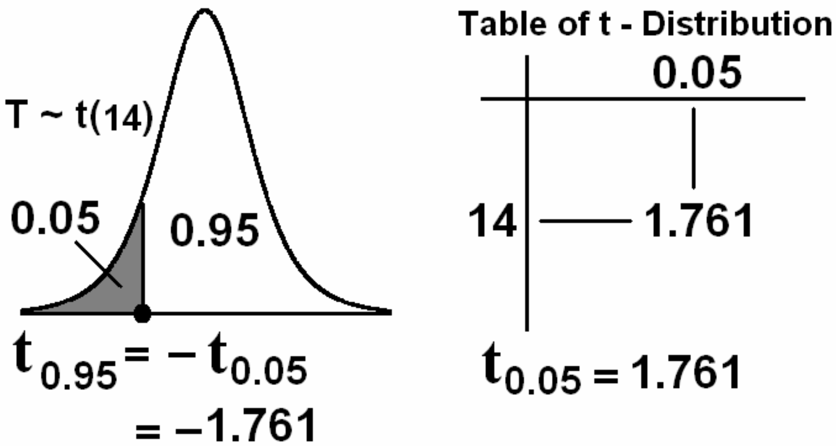
(a) The t-value that leaves an area of 0.95 to the left is

$t_{0.05} = 1.761$



(b) The t-value that leaves an area of 0.95 to the right is

$t_{0.95} = -t_{1-0.95} = -t_{0.05} = -1.761$



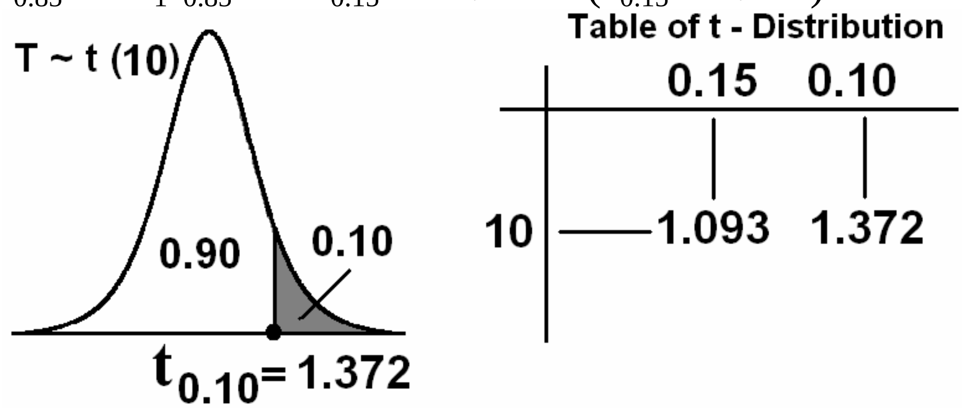
Example:

For $\nu = 10$ degrees of freedom (df), find $t_{0.10}$ and $t_{0.85}$.

Solution:

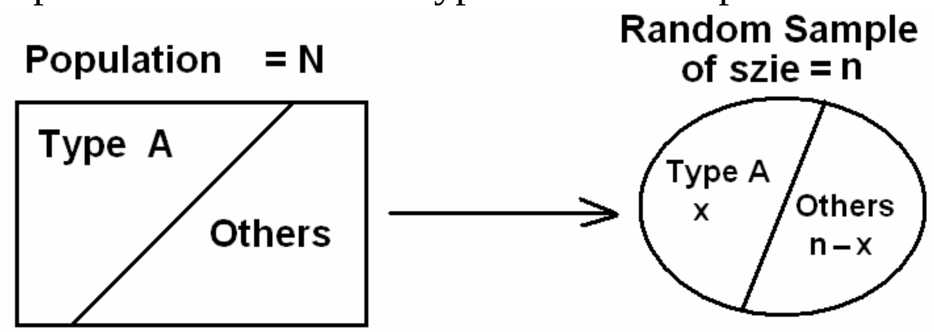
$t_{0.10} = 1.372$

$t_{0.85} = -t_{1-0.85} = -t_{0.15} = -1.093$ ($t_{0.15} = 1.093$)



Sampling Distribution of the Sample Proportion:

Suppose that the size of a population is N. Each element of the population can be classified as type A or non-type A. Let p be the proportion of elements of type A in the population. A random sample of size n is drawn from this population. Let $\hat{p}$ be the proportion of elements of type A in the sample.



Let X = no. of elements of type A in the sample

p =Population Proportion

$$= \frac{no.\ of\ \ elements\ of\ \ type\ \ A\ in\ the\ \ population}{N}$$

$\hat{p}$ = Sample Proportion

$$= \frac{no.\ of\ \ elements\ of\ \ type\ \ A\ in\ the\ sample}{n} = \frac{X}{n}$$

Result:

(1) X ~ Binomial (n, p) {E(X)=np , Var(X)=npq}

(2) $E(\hat{p}) = E(\frac{X}{n}) = p$

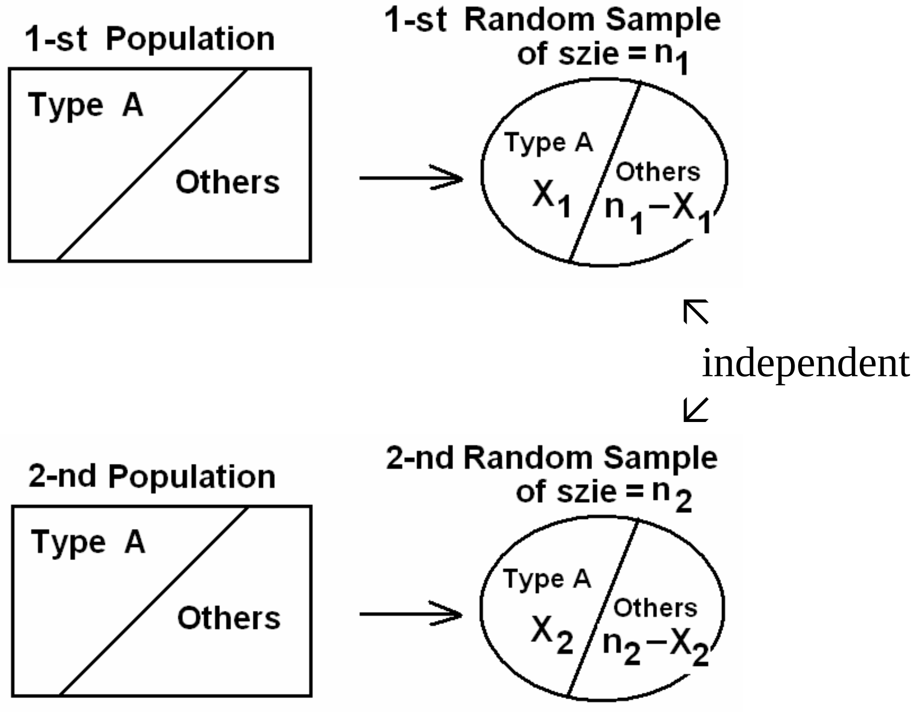
(3) $Var(\hat{p}) = Var(\frac{X}{n}) = \frac{pq}{n}$; $q = 1 - p$

(4) For large n, we have

$$\hat{p} \sim N(p, \sqrt{\frac{pq}{n}}) \quad \text{(Approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \quad \text{(Approximately)}$$

Sampling Distribution of the Difference between Two Proportions:



Suppose that we have two populations:

- p_1 = proportion of the 1-st population.
- p_2 = proportion of the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let X_1 = no. of elements of type A in the 1-st sample.
- Let X_2 = no. of elements of type A in the 2-nd sample.
- $\hat{p}_1 = \frac{X_1}{n_1}$ = proportion of the 1-st sample
- $\hat{p}_2 = \frac{X_2}{n_2}$ = proportion of the 2-nd sample
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.

Result:

(1) $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

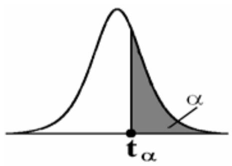
(2) $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1\, q_1}{n_1} + \frac{p_2\, q_2}{n_2} \quad ; q_1 = 1 - p_1,\ q_2 = 1 - p_2$

(3) For large n_1 and n_2, we have

$$\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \sqrt{\frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}}) \quad \text{(Approximately)}$$

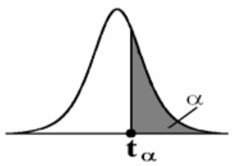
$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}}} \sim \mathrm{N}(0,1) \quad \text{(Approximately)}$$

Critical Values of the t-distribution (t_α)



ν	α						
	0.40	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Critical Values of the t-distribution (t_α)



ν	α						
	0.02	**0.015**	**0.01**	**0.0075**	**0.005**	**0.0025**	**0.0005**
1	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.125	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.291