CHAPTER 7: DISTRIBUTION OF SAMPLE STATISTICS

Sampling from a Population

- 2, 4, 6, 6, 7, 8 sayılarından oluşan bir populasyonumuz olsun
- Bu sayılardan 3 elemanlı bir örneklem (sample) seçebiliriz. Bu elemanlar da 2, 6, 7 olsun.
- Bu 3 sayının ortalaması 5'tir.
- Ama populasyonumuzun ortalaması 5.5'tir.

- Örneklem seçmeye devam edersek

Örneklem	Ortalama
2, 6, 7	5
2, 7, 8	5.7
4, 7, 8	6.33
2, 4, 7	4.33

- Burada 3 elemanlı örneklemlerin ortalamalarının ne kadar değişebileceği (4.33, 5,..., 5.66) hakkında fikir sahibi olduk (distribution of sample means)

- Örneklem kullanmanın (sampling) ve örneklem dağılımını (sampling distribution) bulmanın en önemli yararlarından biri, örneklemin seçildiği populasyonun dağılımı normal olsun ya da olmasın, örneklemin dağılımının normal dağılıma yaklaştığıdır (Central Limit Theorem). Dolayısıyla bir çok test örneklem dağılımı üzerinde uygulanabilir

Sampling Distribution of Sample Means

- We denote population mean with μ, and population variance with σ^2.
- Let's denote a random sample from this population by X, and the unknown elements of X by $X_1, X_2, ..., X_n$.

- *The expectation of sample mean* is defined as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{where } E(\bar{X}) = \mu$$

- Proof:

$$E(\bar{X}) = E(\frac{1}{n}(X_1 + X_2 + ... + X_n)) = \frac{n\mu}{n} = \mu$$

This proof follows the fact that each unknown observation has an expected mean of μ

- *The variance of sample mean* is defined as:

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

- Proof:

$$\begin{aligned}\sigma^2_{\bar{X}} &= Var(\bar{X}) = var(\frac{1}{n}(X_1 + X_2 + ... + X_n)) \\ &= (\frac{1}{n})^2 \sum_{i=1}^{n} \sigma_i^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

This proof follows the fact that the randomly selected observations have zero covariance

- *Central Limit Theorem:* As n becomes large, the distribution of

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

approaches the standard normal regardless of the underlying probability distribution. That is

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- Not 1: Genel olarak örneklem dağılımının normal dağılıma yaklaşması için en az 25 elemanlı olmaları gerektiği gibi bir kabul vardır

- Not 2: İşlediğimiz bu bölümde populasyon ortalama ve varyansının bilindiğini kabul edip örneklem dağılımı hakkında çıkarım yapıyoruz.
  - Gerçekte–ve ileriki bölümlerde işleyeceğimiz üzere– işleyiş bu durumun tam tersi oluyor. Elimizde olan örneklemden populasyon parametreleri hakkınd çıkarım yapılıyor.
    "*Parameters are numbers that describe the properties of entire populations. Statistics are numbers that describe the properties of samples.*"

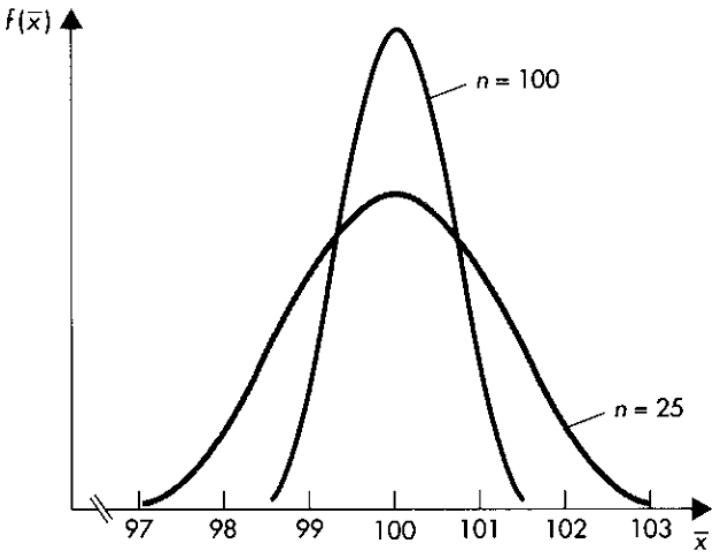
- Örnek: Bir işyerinde çalışanların yıllık maaş artışları ortalaması %12.2, standart sapması %3.6 olan normal bir dağılım göstermektedir. Bu çalışanlarından 9 kişilik bir örneklem alındığında, örneklem ortalamasının %14.4'ten fazla olma ihtimali nedir?

- Verilenler: $\mu = 12.2 \quad \sigma = 3.6 \quad n = 9$

$$\Rightarrow \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{9}} = 1.2$$

$$P(14.4 < X) = P(\frac{14.4 - 12.2}{1.2} < Z)$$
$$= P(1.83 < Z) = 1 - F(1.83) = 0.0336$$

- The standard deviation of the distribution of $\bar{X}$ decreases when sample size, n, increases



- *Law of large numbers:* Central limit theorem states that $\bar{X} \sim N(\mu, \sigma^2/n)$.
- Hence, as n become large, the mean of the samples, $\bar{X}$, converges to the population mean, μ.
- Örnek: Bir populasyonun dağılımının (normal olsun ya da olmasın) ortalaması 5, standart sapması da 2 olsun. Bu populasyondan seçilen çeşitli bütüklükteki örneklemlerin (n=10, n=25, n=50) dağılıma bakalım

– Burada bize verilen; $\mu = 5$ ve $\sigma = 2$. Şimdi farklı elemanlı örneklemlerin dağılımlarına bakalım

– If $n = 10$, $\bar{X} \sim N(\mu, \sigma^2/n)$

$* \Rightarrow \quad \bar{X} \sim N(5, 2^2/10) \quad \Rightarrow \quad \bar{X} \sim N(5, 0.4)$

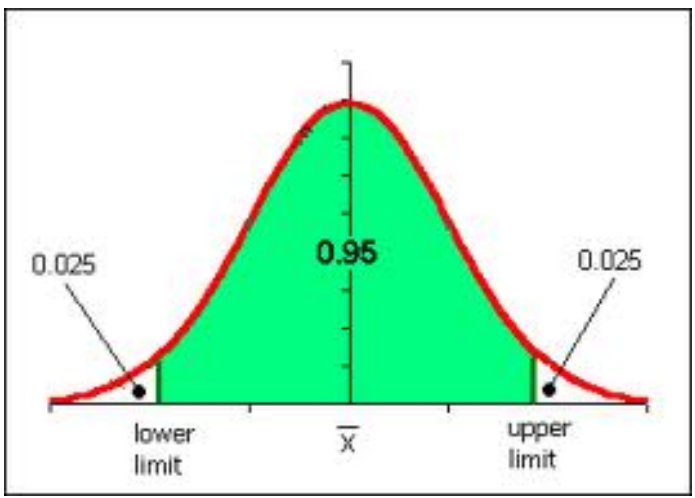
– If $n = 25$, $\bar{X} \sim N(\mu, \sigma^2/n)$

$* \Rightarrow \quad \bar{X} \sim N(5, 2^2/25) \quad \Rightarrow \quad \bar{X} \sim N(5, 0.16)$

– If $n = 100$, $\bar{X} \sim N(, \sigma^2/n)$

$* \Rightarrow \quad \bar{X} \sim N(5, 2^2/100) \quad \Rightarrow \quad \bar{X} \sim N(5, 0.04)$

- Örneklemin eleman sayısı arttıkça örneklem ortalamasının varyasyonunun düştüğünü görmenin başka bir yolu ise bu örneklem dağılımları için data'nın %95'ini içeren güven aralığına bakmak olabilir



- Standart normal tablosundan datanın 0.975'ini içeren Z değerini bulabiliriz; bu da 1.96'dır.
- Eğer $n = 10$ ise, datanın %95'ini içeren aralık

$$[5 - 1.96 * 2/\sqrt{10} < \bar{X} < 5 + 1.96 * 2/\sqrt{10}]$$
$$\Rightarrow \quad [4.2 < X < 6.3]$$

- Eğer $n = 25$ ise,

$$[5 - 1.96 * 2/\sqrt{25} < \bar{X} < 5 + 1.96 * 2/\sqrt{25}]$$
$$\Rightarrow \quad [4.2 < X < 5.8]$$

- Eğer $n = 100$ ise,

$$[5 - 1.96 * 2/\sqrt{100} < \bar{X} < 5 + 1.96 * 2/\sqrt{100}]$$
$$\Rightarrow \quad [4.6 < X < 5.4]$$

  - Örneklemlerin eleman sayısı arttıkça, örneklem ortalaması giderek populasyon ortalaması etrafında daha sıkı bir şekilde dağılmaktadır

Sampling Distribution of Sample Proportions

***Bernoulli distribution**

- Suppose there is a random experiment takes the value 1 with probability P and the value 0 with probability $(1-P)$. And suppose this experiment is conducted only once (a single trial). Then we say that the resulting random variable X has a Bernoulli distribution where

$$f(x; P) = P^x(1-P)^{1-x} \quad \text{for } x = 0, 1$$

or

$$f(x;P) = \left\{ \begin{array}{cc} P & \text{if } x = 1 \\ 1-P & \text{if } x = 0 \end{array} \right\}$$

- Then the mean in

$$E(X) = 1(P) + 0(1-P) = P$$

- The variance is

$$\sigma_X^2 = (0-P)^2(1-P) + (1-P)^2(P) = P(1-P)$$

***The Binomial Distribution**

- Suppose there is a random experiment takes the value 1 with probability P and the value 0 with probability $(1-P)$. And suppose this experiment is conducted n times . Then we say that the total number of successes X is random variable that follows the Binomial Distribution

$$P(x; n, P) = \begin{pmatrix} n \\ x \end{pmatrix} P^x (1-P)^{n-x} \quad \text{for x=0,1, 2, ...,n}$$

- Ex: Havaya atılan bir madeni para sonucunda Yazı gelmesini başarı olarak adlandıralım ve para 3 kez havaya atıldığında toplam kaç kez Yazı gelebileceğinin ihtimallerine bakalım. Sırasıyla 0, 1, 2 veya 3 kez gelebilir:

$$P(0;3,0.8) = \left(\begin{array}{c} 3 \\ 0 \end{array} \right) 0.5^0(1-0.5)^3 = 0.125$$

$$P(1;3,0.8) = \left(\begin{array}{c} 3 \\ 1 \end{array} \right) 0.5^1(1-0.5)^2 = 0.375$$

$$P(2;3,0.8) = \left(\begin{array}{c} 3 \\ 2 \end{array} \right) 0.5^2(1-0.5)^1 = 0.375$$

$$P(3;3,0.8) = \left(\begin{array}{c} 3 \\ 3 \end{array} \right) 0.5^3(1-0.5)^0 = 0.125$$

which sums up to 1.

- If each trial is called X_i, and if trials are repeated n times, the total number of successes is

$$X = X_1 + X_2 + ...X_N$$

- The mean of X is

$$E(X) = E(X_1 + X_2 + ...X_N) = nP$$

- The variance of the binomial distribution is

$$\begin{aligned} Var(X) &= Var(X_1 + X_2 + ...X_N) \\ &= nVar(Xi) = nP(1-P) \end{aligned}$$

Sample Proportion

- Let X be the number of successes in a binomial sample of n observations and P probability of success for each of these observations.
- Then the proportion of successes (*başarı oranı*)

$$\hat{p}_x = \frac{X}{n}$$

in the sample is called the *sample proportion*

- The mean of $\hat{p}_x = \dfrac{X}{n}$ is

$$E(\hat{p}_x) = E(\frac{X}{n}) = \frac{E(X)}{n} = \frac{nP}{n} = P$$

- The variance of $\hat{p}_x$ is

$$Var(\hat{p}_x) = Var(\frac{X}{n}) = \frac{Var(X)}{n^2}$$
$$= \frac{nP(1-P)}{n^2} = \frac{P(1-P)}{n}$$

- The standard deviation of $\hat{p}_x$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$$

- If the sample size is large, then the following statistics is distributed approximately as standard normal

$$Z = \frac{\hat{p}_x - E(\hat{p}_x)}{\sigma_{\hat{p}}} = \frac{\hat{p}_x - P}{\sigma_{\hat{p}}}$$

- Öğrenciler arasında yapılan bir araştırmaya göre ahlak dersinin öğrencilerin davranışları üzerinde olumlu etki yarattığına inananların oranı %43'tür. Bu öğrenciler içerisinden rastgele seçilen 80 kişilik bir gruptakilerin yarısından fazlasının bu görüşte olma olasılığı nedir?

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.43 * 0.57}{80}} = 0.055$$

– Aradığımız ihtimal

$$P(.5 < \hat{p}_x) = P(\frac{.5 - P}{\sigma_{\hat{p}}} < \frac{\hat{p}_x - P}{\sigma_{\hat{p}}})$$

$$= P(\frac{.5 - .43}{.055} < Z) = P(1.27 < Z)$$

$$= 1 - F_Z(1.27) = 1 - .8980 = .1020$$

- Örnek: Bir bölgedeki insanların %20'sinin A partisine oy veriyor olduğunu düşünelim. Buradan seçilecek 270 tane kişinin %16 ila %24 arasında olan oranının A partisine oy verme olasılığı kaçtır?
  - Burada bize verilenler $P = 0.2$ ve $n = 270$. Dolayısıyla

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.2 * 0.8}{270}} = 0.024$$

– Aradığımız ihtimal ise

$$P(.16 < \hat{p}_x < .24)$$

$$= P(\frac{.16 - 0.2}{0.024} < Z < \frac{.24 - 0.2}{0.024})$$

$$= P(-1.67 < Z < 1.67) = F(1.67) - F(-1.67)$$

$$= F(1.67) - [1 - F(1.67)]$$

$$= .9525 - (1 - .9525) = .9050$$

Sampling Distribution of Sample Variances

- Let the population variance is given as

$$\sigma^2 = E[(X - \mu)^2]$$

- And the sample variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

- To analyze whether the sample variance is a good approximation to the population variance, we will need both expected value of the sample variance, and also its variance

- Let's denote the random sample by X, and its unknown elements by $X_1, X_2, ..., X_n$. The *sample variance* is defined as follows

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

- We already know that

$$E(s^2) = \sigma^2$$

- If we rearrange the last two equations, we obtain the following random variable

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

which, under the assumption of normally distributes population, has a χ^2 (chi-square) distribution with $n-1$ degrees of freedom

- We denote the χ^2 distribution with ν degrees of freedom by χ^2_ν. And the mean and variance of this distribution are

$$E(\chi^2_\nu) = \nu \qquad \text{and} \qquad Var(\chi^2_\nu) = 2\nu$$

- As our random variable $(n-1)s^2/\sigma^2$ has a distributed with $\chi^2_{(n-1)}$, its mean and the variance can be written as

$$E[\frac{(n-1)s^2}{\sigma^2}] = (n-1)$$

$$Var[\frac{(n-1)s^2}{\sigma^2}] = 2(n-1)$$

– Using

$$E[\frac{(n-1)s^2}{\sigma^2}] = (n-1)$$

we obtain

$$E(s^2) = \sigma^2$$

– Using

$$Var[\frac{(n-1)s^2}{\sigma^2}] = 2(n-1)$$

we obtain

$$Var(s^2) = \frac{2\sigma^4}{(n-1)}$$

- Örnek: Üretilen bir malın dayanım süresi normal bir dağılıma sahip olup, 3.6 oranında standart sapmaya sahiptir. Bu mallardan 4 elemanlı rastgele bir örneklem seçilirse, bu örneklemin dayanım süresinin varyasyonunun 30'dan büyük olma ihtimali kaçtır?
  - Cevap

$$P(s^2 > 30) = P(\frac{(n-1)s^2}{\sigma^2} > \frac{30(n-1)}{\sigma^2})$$

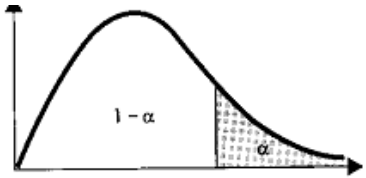
$$= P(\chi^2_3 > \frac{30 * 3}{3.6^2}) = P(\chi^2_3 > 6.94)$$

- Chi-square tablosunda 3 serbestlik dereceli bir dağılım için 6.94'ü kapsayacak iki değer 6.25 ve 7.81'dir. Bunların karşılığı ise

$$P(\chi_3^2 > 6.25) = .1 \qquad P(\chi_3^2 > 7.81) = .05$$

dolayısıyla tekrar aradığımız ihtimal

$$0.05 < P(s^2 > 30) < .10$$



For selected probabilities α, the table shows the values $\chi^2_{\nu,\alpha}$ such that $\alpha = P(\chi_\nu^2 > \chi^2_{\nu,\alpha})$, where χ_ν^2 is a chi-square random variable with ν degrees of freedom. For example, the probability is .100 that a chi-square random variable with 10 degrees of freedom is greater than 15.99.

ν	α									
	.995	.990	.975	.950	.900	.100	.050	.025	.010	.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95

- Örnek: Cips paketlerinin ağırlığının normal dağılıma sahip olduğu varsayılırsa, 20 tane rastgele seçilen paket için aşağıdaki ihtimali sağlayan K'yı bulalım

$$P(\frac{s^2}{\sigma^2} < K) = 0.05$$

*Burada rneklem varyasyonunun popülasyon varyasyonuna oranının K veya daha küçük olma ihtimali %5 olarak tanımlanıyor.

– Aradığımız ihtimal aşağıdaki şekilde yazılabilir

$$0.05 = P(\frac{s^2}{\sigma^2} < K) = P(\frac{(n-1)s^2}{\sigma^2} < (n-1)K)$$

$$0.05 = P(\chi^2_{(n-1)} < (n-1)K)$$

$$0.05 = P(\chi^2_{19} < 19K)$$

Chi-square tablosundan 19 serbestlik dereceli bir chi-square dağılımı için

$$19K = 10.12 \quad \Rightarrow \quad K = 0.533$$

Örneklem varyasyonunun popülasyon varyasyonunun %53'ünden küçük olma ihtimali %5'dir