BENZETİM

DR.ÖĞR. ÜYESİ HACI MEHMET ALAKAŞ

GİRDİ MODELLEME

Bir kuyruk sistemi simülasyonunda tipik girdi verisi varışlar arası zaman ve servis zamanlarının dağılımıdır.

Bir envanter sistemi simülasyonu için girdi verisi talep ve temin zamanlarının dağılımlarını içerir.

Bir güvenilirlik sistemi simülasyonu için bir bileşen arızalanana kadar ki geçen sürenin dağılımı girdi verisine örnek verilebilir.

Şu ana kadar ki yapılan örneklerde ve uygulamalarda uygun dağılımlar belirlenmiş ve hazır olarak bize verilmişti. Ancak gerçek hayata ilişkin simülasyon uygulamalarında girdi verisi için uygun dağılımların belirlenmesi zaman ve kaynak gereksinimi açısından zor bir iştir.

Girdilere ilişkin yapılan hatalı modeller çıktılarında hatalı olmasına neden olacak ve sistemlere ilişkin yanlış öneriler anlamına gelecektir.

Girdi verisine ait kullanışlı bir model geliştirilmesinde dört adım vardır:

1. **İlgilenilen gerçek sistemden veriler toplamak.** Bu genellikle ciddi bir zaman ve kaynak gereksinimine ihtiyaç duymaktadır. Maalesef, bazı durumlarda veri toplamak mümkün değildir (örneğin, zaman çok kısıtlıysa, girdi prosesi hali hazırda yoksa, veya yasalar veya kurallar veri toplanmasını yasaklıyorsa). Veri mevcut değilken, uzman görüşü veya süreç bilgisi uygun tahminler yapmak için kullanılmalıdır.
2. **Girdi prosesini temsil etmek üzere bir olasılık dağılımının belirlenmesi.** Veri mevcut iken, bu adım genellikle verinin bir frekans (sıklık) dağılımı veya histogramının geliştirilmesiyle başlar. Frekans dağılımına ve prosesin yapısal bilgisine dayalı olarak bir dağılım ailesi seçilir. Neyse ki olasılık ve istatistik modellerinin anlatıldığı bölümdeki gibi bir çok iyi bilinen dağılım pratikte iyi yaklaşımlar sağlamaktadır.

3. **Dağılım ailesinin spesifik bir durumunu belirleyecek parametrelerin seçimi**. Veriler mevcut iken bu parametreler veriden tahmin edilebilir.
4. **Seçilen dağılımın ve ilgili parametrelerin uyum-iyiliği (goodness-of-fit) için değerlendirilmesi.** Uyum-iyiliği informal olarak grafiksel tekniklerle ve formal olarak istatistiksel testlerle değerlendirilebilir. Ki-Kare ve Kolmogorov-Smirnov testleri standart uyum-iyiliği testleridir. Analist seçilen dağılımın verinin uygun bir yaklaşımı olduğuna ilişkin olarak tatmin olmadıysa ikinci adıma geri döner ve farklı bir dağılım ailesi seçerek prosedürü tekrarlar. Eğer bu prosedürün birkaç kez tekrarlanması neticesinde varsayılan dağılım formu ile toplanan veri arasında bir uygunluk elde edilmezse dağılımın ampirik formu kullanılabilir.

VERİ TOPLAMA

Veri toplama gerçek bir hayat probleminin çözülmesindeki en büyük görevlerden birisidir.

Simülasyonda ise en önemli ve en zor problemlerden biridir. Veriler hali hazırda mevcut iken dahi nadiren simülasyon girdi modeli için direkt olarak kullanılabilir bir formdalardır.

Model yapısı geçerli olsa dahi veriler hatalı şekilde toplandıysa, uygun olmayan bir şekilde analiz edildiyse veya ortamın iyi bir temsili değilse simülasyon çıktıları yanıltıcı olacak ve karar verme sürecinde veya politika geliştirmede kullanıldığında muhtemelen zarar verici ve maliyetli olacaktır.

ÖRNEK(ÇAMAŞIRHANE)

Simülasyon dersini alan iki öğrencinin hali hazırda çalışan bir sistemin operasyonunu simüle etmek için bazı görevleri vardır. Oldukça basit çalışıyor gibi görülen bu sistemlerden biri 10 çamaşır makinası ve 6 kurutma makinası ile çalışan bir self-servis çamaşırhanedir.

Ancak problemin veri toplama kısmı hızlı bir şekilde devasa bir boyuta ulaşmıştır. Varışlar arası zaman dağılımı homojen değildi yani dağılım haftanın farklı günlerinde ve günün farklı zamanlarında değişiyordu. Çamaşırhane haftanın 7 günü ve günde 16 saat açıktı (veya haftada 112 saat). Mevcut sınırlı kaynaklarla (bu iki öğrenci 4 farklı ders daha alıyor) ve zaman kısıtı altında (simülasyonun 4-haftalık bir zaman periyodunda tamamlanması gerekiyordu) çamaşırhanenin tüm operasyonlarını kapsamak mümkün görünmüyordu. Ek olarak bir hafta boyunca varışlar arası zamanın dağılımı bir sonraki hafta boyunca farklılaşabiliyordu. Uzlaşık bir çözüm olarak zaman örneklemi seçildi ve varışlar arası zaman dağılımı belirlendi ve varış oranına bağlı olarak (belki de uygun olmayan bir şekilde) “yüksek”, “orta” ve “düşük” olarak sınıflandırıldı.

Servis zamanı dağılımları da bir çok açıdan zor bir problem olarak karşılarına çıkmıştı. Çeşitli servis kombinasyonları isteyen müşterilerin oranının gözlemlenmesi ve kaydedilmesi gerekiyordu.

En basit durum bir çamaşır makinasının ardından bir kurutucu isteyen müşteriydi. Ancak bir müşteri bir çamaşır makinasından sonra bir kurutma makinası, ya da sadece bir kurutma makinası gibi seçeneklerde bulunabilir.

Müşteriler numaralandırılmış makine kullandıklarından onları kişisel karakterleriyle hatırlamak yerine bu referansı kullanarak onları takip etmek mümkündü.

Herhangi bir müşteri için çamaşır makinası talebi ile kurutma makinası talebi arasında bir bağımlılık olduğundan çamaşır ve kurutma makinaları için servis zamanlarını bağımsız değişkenler gibi birbirinden ayrık olarak ele almak uygun olmayacaktı.

Bazı müşteriler çamaşırlarının yıkanmasının veya kurutulmasının tamamlanmasını sabırla bekledikleri ve sonrasında hemen makinayı boşaltmaktaydılar. Diğerleri ise çamaşırlarını orada bırakmakta makine kendi çamaşırlarının yıkanması için gereken süre tamamlandıktan bir süre sonra dönmekteydiler. Yoğun bir zaman diliminde gelmeyen müşterinin çamaşırlarını makineden alıp bir sepetin içerisine çıkarmaktadır. Servis tamamlanma zamanı çamaşırlar makineden çıkarıldığı zaman olarak tanımlanmıştır.

Ayrıca makinalar zaman zaman arızalanmaktadır. Arızada geçirilen süre birkaç dakikadan (yönetici makineyi tamir ettiği takdirde) birkaç güne kadar (Cuma akşamı oluşan ve çamaşırhanede olmayan bir parçaya ihtiyaç duyulan bir arıza bir sonraki pazartesiye kadar tamir edilemez) değişebilmektedir. Kısa süreli tamir tamamlanma zamanları öğrenciler tarafından kaydedilmiştir. Uzun-dönemli tamir tamamlanma zamanları yönetici tarafından tahmin edilmiştir. Dolayısıyla arızalanmalar simülasyonun bir parçasıdır.

Veri toplama esnasındaki gerçek tecrübelerden bir çok şey öğrenilebilir.

Aşağıdaki tavsiyeler veri toplama işlemini kolaylaştırabilir.

1. Planlama sayesinde zamandan ciddi tasarruflar yapılabilir. Bu sistem üzerinde ön gözlemlerle başlayabilir. Ön gözlem esnasında veri toplamaya çalışın. Bu amaç için formlar geliştirin. Bu formlar yüksek ihtimalle gerçek veri toplama işlemi başlamadan önce birkaç kez değiştirilecektir. Olağan dışı durumları gözleyin ve bunları nasıl ele alacağınızı düşünün. Mümkünse sistemin video kaydını alın bu kaydı daha sonra inceleyerek gerekli veriyi buradan çıkartın. Veriler otomatik olarak toplanıyor olsa dahi (örneğin bilgisayar desteğiyle) uygun verilerin mevcut olup olmadığından emin olmak için planlama yapmak önemlidir. Veriler zaten başka biri tarafından toplandıysa verileri kullanılabilir bir formata dönüştürmek için oldukça fazla zaman ayırmanız gerekebilir.

2. Veriler bir taraftan toplanmaya devam edilirken verileri analiz etmeye çalışın. Toplanan verilerin simülasyona girdi için gerekli dağılımları sağlamada yeterli olup olmadığını belirleyin. Toplanan herhangi bir verinin simülasyon için gereksiz olup olmadığını belirleyin. Lüzumsuz veri toplamanın bir anlamı yoktur.
3. Homojen veri setlerini birleştirmeye çalışın. Ardışık zaman periyotlarında ve ardışık günlerin aynı zaman dilimlerinde homojenlik için verileri kontrol edin. Örneğin, verilerin homojenliğinin saat 14:00-15:00 arası ve 15:00-16:00 arası için kontrol edin ve Perşembe ve Cuma günleri için saat verileri 14:00- 15:00 arasında kontrol edin. Homojenliği kontrol ederken başlangıçta yapılacak test dağılımların ortalamalarının aynı olup olmadığını görmektir. İki örneklem t-testi bu amaç için kullanılabilir.

4. Veri sansürleme ihtimaline karşı dikkatli olun. İlgilenilen veri miktarı tamamıyla gözlemlenmemiş olabilir. Bu problem çoğunlukla analist belirli bir süreci tamamlamak için gerekli zamanla ilgilendiğinde oluşmaktadır (örneğin, bir ürünün üretimi, bir hastanın tedavisi gibi). Ancak süreç gözlemin başladığı zamandan daha önce başlar veya daha sonra tamamlanır. Sansürleme veri örnekleminin dışında kalmış uzun işlem zamanları anlamına gelebilir.
5. İki değişken arasında bir ilişkinin olup olmadığını belirlemek için serpilme diyagramı (scatter diagram) oluşturun. Bazen serpilme diyagramına göz ucuyla bakmak ilgilenilen iki değişken arasında ilişki bulunup bulunmadığını bize gösterecektir.

6. Bağımsız gibi görünen gözlem dizisinin içinde bir oto korelasyon bulunma ihtimalini düşünün. Oto korelasyon ardışık zaman periyotları veya ardışık müşteriler içinde oluşabilir. Örneğin, örneğin i. müşterinin servis zamanı $(i+n)$. müşterinin servis zamanıyla ilişkili olabilir.

7. Girdi verisi ile çıktı veya performans verisi arasındaki farkı aklınızda tutun ve girdi verisi topladığınızdan emin olun. Girdi verisi tipik olarak geniş çapta sistemin kontrolünün ötesinde belirsiz sayıları/miktarları temsil eder ve sistemi iyileştirmek için yapılan müdahalelerle değişmeyecektir. Diğer yandan çıktı verisi girdi verilerine bağımlı sistem performansını temsil eder. Kuyruk simülasyonunda müşteri varış zamanları daima girdiler iken müşteri gecikmeleri/beklemeleri bir çıktıdır. Performans verileri model geçerliliği için kullanışlıdır.

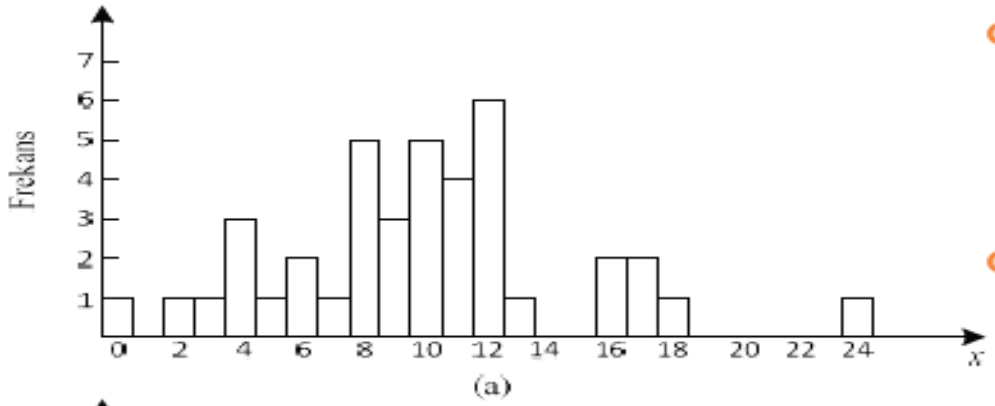
VERİLERLE DAĞILIMIN BELİRLENMESİ

- Bu bölümde veriler mevcut iken girdi dağılım ailelerinin seçimine ilişkin metotlardan bahsedilecektir. Bir aile içerisindeki spesifik bir dağılım parametrelerinin tahmin edilmesiyle belirlenir.

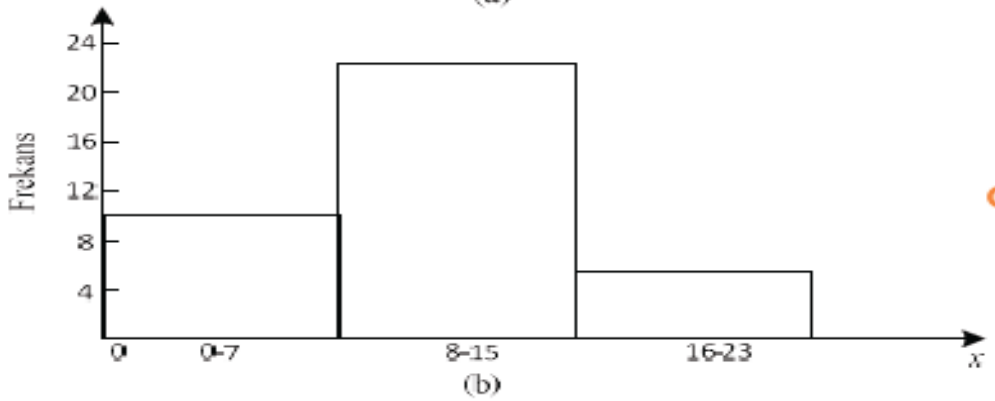
Histogramlar

Frekans dağılımı veya histogram dağılımın şeklini belirlemede kullanışlıdır. Bir histogram aşağıdaki şekilde oluşturulur:

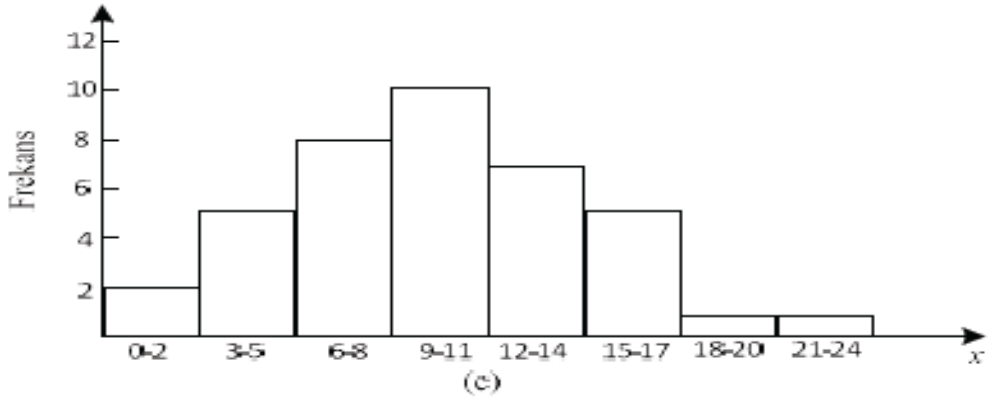
1. Verileri aralıklara bölün (aralıklar genellikle eşit genişlikte olur ancak frekansların yükseklikleri ayarlanacaksa eşit olmayan genişlikler de kullanılabilir).
2. Yatay ekseni seçilen aralıklara uygun şekilde etiketleyin.
3. Her bir aralıktaki oluşumların frekanslarını bulun.
4. Toplam gözlem değerlerini her bir aralık için gösterecek şekilde düşey ekseni etiketleyin.
5. Düşey eksen üzerindeki frekansları gösteriniz.



(a)



(b)



(c)

- Sınıf aralıklarını sayısı gözlem sayısına ve verideki serpilme/yayılmaya bağlıdır.
- Araştırmacılar pratikte sınıf sayısının toplam gözlem sayısının yaklaşık olarak karekökünü eşit olarak seçilmesinin iyi sonuçlar verdiğini belirtmektedir.
- Eğer aralıklar çok genişse histogram çok kaba (coarse) olacak şekli ve diğer detaylar iyi bir şekilde görülmeyecektir. Eğer aralıklar çok darsa histogram çok tırtıklı (ragged) olacak ve veriyi düzleştiremeyecektir. Aynı veriyi kullanarak oluşturulmuş tırtıklı, kaba ve uygun histogramlar yukarıdaki şekilde gösterilmektedir

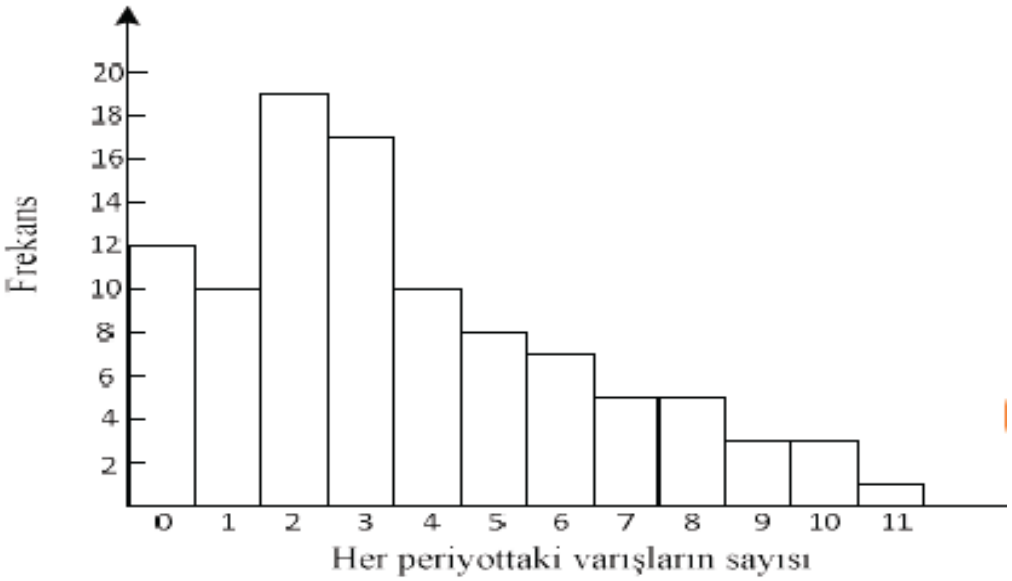
- Sürekli veriler için histogram bir teorik dağılımın olasılık yoğunluk fonksiyonuna uymaktadır. Eğer veriler sürekli ise her bir sınıf aralığı frekansının merkez noktasından geçen bir çizgi çizilir olasılık yoğunluk fonksiyonunun şekline benzemelidir.

- Çok fazla sayıda veri noktasının bulunduğu kesikli veriler için histogramlar, veri aralığındaki her değer için bir hücreye sahip olmalıdır. Ancak birkaç veri noktası varsa histogramın bozuk yapısını ortadan kaldırmak için komşu hücreleri birleştirmek gerekebilir. Eğer histogram kesikli verilerle ilişkili ise şekli bir olasılık kütle fonksiyonuna benzeyecektir.

ÖRNEK: (KESIKLI VERI)

- Bir kavşağın kuzey-batı tarafından gelen araçların sayısı sabah 07:00 ile 07:05 saatleri arasında 5 dakikalık süreyle 20 hafta boyunca haftada 5 işgünü boyunca gözlemlenmiştir. Elde edilen veriler aşağıdaki tabloda verilmiştir. Tablodaki ilk giriş 5-dakikalık bu zaman diliminde 12 defa hiç araç gelmediği anlamına gelmektedir.
- Otomobillerin sayısı kesikli veri olduğu için ve veri sayısı da çok olduğundan veri aralığı içinde her bir muhtemel için bir hücreyi kullanabiliriz. Elde edilen histogram aşağıdaki gibidir:

Her periyottaki varışlar	Frekans	Her periyottaki varışlar	Frekans
0	12	6	7
1	10	7	5
2	19	8	5
3	17	9	3
4	10	10	3
5	8	11	1

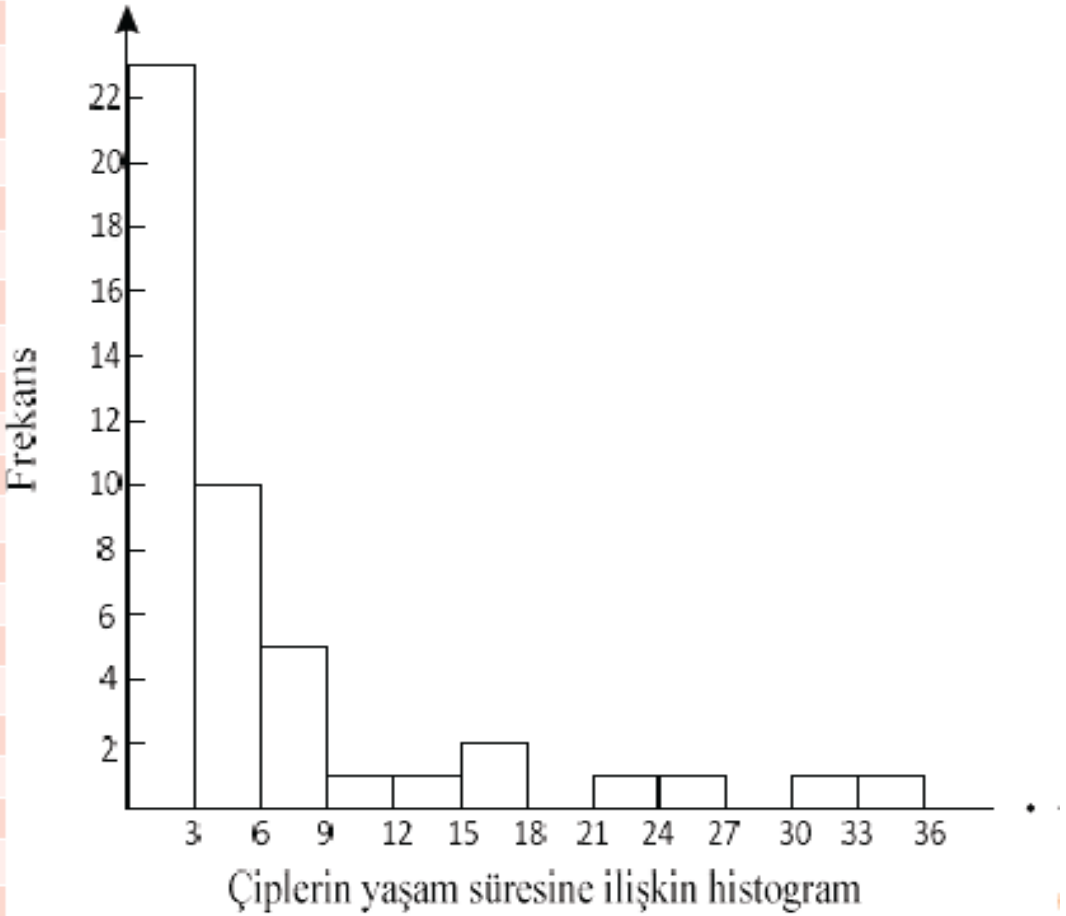


ÖRNEK: (SÜREKLI VERI)

- Normal voltajın 1.5 katı verilerek elektronik çiplerden oluşan bir rassal örnekleme ömür testleri yapılmış ve yaşam süreleri (ya da bozulana kadar geçen süreler) gün cinsinden kaydedilmiştir:

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562

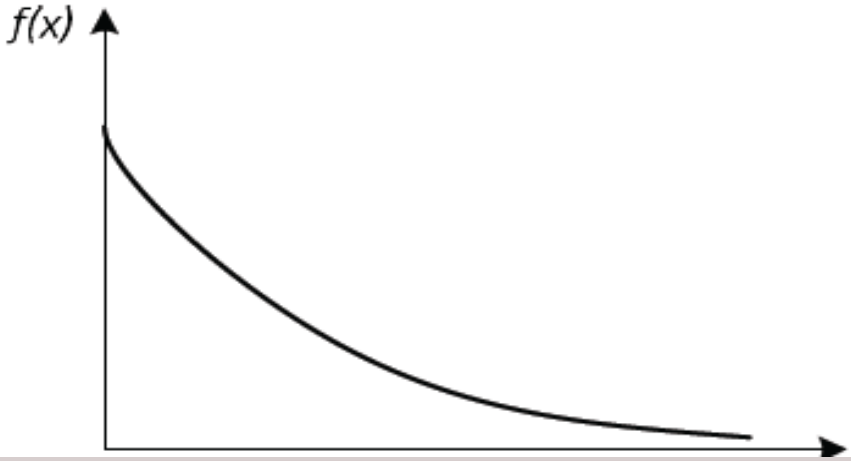
Çip Ömrü (Gün)	Frekans
$0 \leq x_j < 3$	23
$3 \leq x_j < 6$	10
$6 \leq x_j < 9$	5
$9 \leq x_j < 12$	1
$12 \leq x_j < 15$	1
$15 \leq x_j < 18$	2
$18 \leq x_j < 21$	0
$21 \leq x_j < 24$	1
$24 \leq x_j < 27$	1
$27 \leq x_j < 30$	0
$30 \leq x_j < 33$	1
$33 \leq x_j < 36$	1
.	.
.	.
.	.
$42 \leq x_j < 45$	1
.	.
.	.
.	.
$57 \leq x_j < 60$	1
.	.
.	.
.	.
$78 \leq x_j < 81$	1
.	.
.	.
.	.
$144 \leq x_j < 147$	1



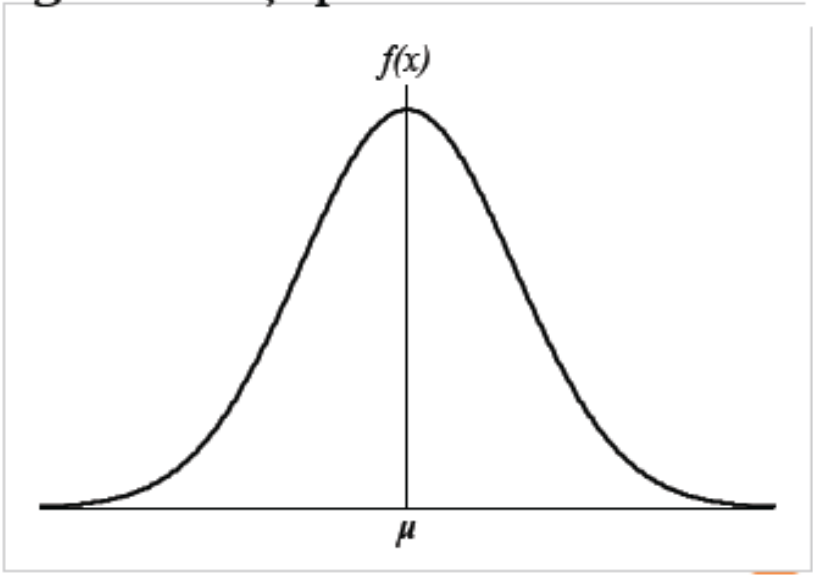
Çiplerin yaşam süresine ilişkin histogram

DAĞILIM AILESININ SEÇIMI

- Bir histogram hazırlamanın nedeni bilinen bir oyf veya okf ortaya çıkarmaktır. Bir dağılım ailesi belirlemek histogramda çıkan şekle dayalı olarak seçilir. Böylelikle eğer varışlar-arası zamana ilişkin veriler toplandıysa ve histogram aşağıdaki şekle benziyorsa üstel dağılımdan şüphelenebiliriz.



- Benzer şekilde nakliye paletlerinin ağırlığına ilişkin ölçümler yapıldıysa ve histogram ortalama simetrik ve şekli aşağıdaki gibiyse bu sefer normal dağılımdan şüphelenebiliriz.



DAĞILIM AİLESİNİN SEÇİMİ

- Çoğu belirli bir fiziksel modele uyan yüzlerce olasılık dağılımı geliştirilmiştir. Dağılımları seçimine yardımcı olacak şeylerden biri dağılımların hangi fiziksel tabana oturtulduğudur. Aşağıda buna ilişkin bazı örnekler verilmiştir:
- **Binom dağılımı** her bir deneyde başarı şansı sabit p olan n adet deney içerisinde başarılı olan deneylerin sayısını modeller; örneğin n adet bilgisayar çipinden oluşan bir partide kusurlu olan çiplerin sayısı.
- **Negatif Binom dağılımı** (geometrik dağılımı da dahil) k tane başarı elde edilinceye kadar ki yapılması gereken deney sayısını modeller; örneğin 4 adet kusurlu çipi bulana kadar ki kontrol etmemiz gereken çip sayısı.
- **Poisson dağılımı** belirli bir sabit zaman veya aralıkta oluşan bağımsız olayların sayısını modeller; örneğin 1 saatlik zaman diliminde bir süpermarkete gelen müşterilerin sayısı veya bir metal levhanın 30 m2'lik alana sahip kısmında bulunan hata sayısı.

- **Normal dağılım** alt süreçlerinin sayılarının toplamı olarak düşünülebilecek bir süreci modeller; örneğin her bir montaj operasyonunun tamamlanması için gerekli zamanları toplamı olan bir parçanın montajı için gerekli zaman. Normal dağılım proses zamanları için mümkün olmayan negatif sayıları da kabul etmektedir.
- **Lognormal dağılım** birçok alt sürecin çarpımı olarak düşünülecek bir sürecin dağılımını modeller; örneğin faiz oranı bileşik (compound) iken bir yatırımın geri dönüş oranı belirli bir periyot sayısı boyunca geri dönüşlerin çarpımıdır.
- **Üstel dağılım** hafızası olmayan bağımsız olaylar veya proses zamanları (bir süreç tamamlanmadan önce ne kadar zamanın geçtiği bilmek o sürecin tamamlanması için ne kadar daha ek süre gerektiği ile ilgili hiçbir bilgi vermez) arasındaki zamanı modeller. Örneğin birbirlerinden bağımsız olarak hareket eden büyük bir müşteri kitlesinin varışları arasındaki zaman. Üstel dağılım oldukça değişken bir dağılımdır ve matematiksel olarak kolay modellere neden olduğundan oldukça sık kullanılmaktadır.

- **Gamma dağılımı** pozitif rassal değişkenlerin modellenmesinde kullanılan oldukça esnek bir dağılımdır. Gamma bir sabit ekleyerek o'dan uzaklaştırılabilir.

- **Beta dağılımı** sınırlandırılmış (sabit alt ve üst limitler) rassal değişkenlerin modellenmesinde kullanılmaktadır. Beta bir sabit eklenerek o'dan uzaklaştırılabilir ve bir sabitle çarpılarak $[0, 1]$ aralığından daha geniş aralığa sahip olabilir.

- **Erlang dağılımı** üstel dağılmış birkaç sürecin toplamı olarak görülebilecek süreçlerin modellenmesinde kullanılır. Örneğin, bir bilgisayar ağı bir tane bilgisayar ve iki adet yedek bilgisayar arızalandığında çökmekte ise her bir bilgisayar arızalanana kadar ki geçen zaman üstel dağılımla modelleniyorsa Erlang dağılımı bilgisayar ağı çökene kadar ki geçen zamanın modellenmesinde kullanılır. Erlang dağılımı Gamma'nın özel bir halidir.

- **Weibull dağılımı** herhangi bir cihazın arızalanana kadar ki geçen sürenin modellenmesinde kullanılmaktadır. üstel dağılım Weibull dağılımının özel bir halidir.

- **Kesikli veya Sürekli Düzgün dağılım** tüm çıktıların gerçekleşme ihtimali eşit olduğundan tam belirsizliği modeller. Bu dağılım verinin olmadığı durumda oldukça fazla kullanılmaktadır.

- **Üçgen dağılım** ise dağılımın en küçük, en çok tekrar eden ve en büyük değerleri olduğunda bir süreci modeller.

- **Ampirik dağılım** toplanmış gerçek veriden yeniden örneklem alınması işini yapar. Genellikle herhangi bir teorik dağılım uygun bulunmadığında kullanılır.

- Bir dağılım seçerken sürecin fiziksel karakteristiğini göz ardı etmeyin.

- Sürecin kendi doğası kesikli mi yoksa sürekli değerler mi alıyor? Veriye bağlı olmayan bu bilgi hangi dağılım ailesinin seçileceği konusundaki alternatifleri azaltmamıza yardımcı olur.

- Ayrıca şu gerçek hiçbir zaman unutulmamalıdır: herhangi bir stokastik girdi prosesi için "gerçek" bir dağılım yoktur.

- Bir girdi modeli gerçeğe bir yaklaşım sunar o halde amaç simülasyon deneyinden kullanışlı sonuçlar üreten bir yaklaşık sonuç elde etmektir.

PARAMETRE TAHMİNİ

- Dağılım ailesi seçildikten sonra bir sonraki aşama dağılımın parametrelerinin tahmin edilmesidir.
- Birçok dağılım için tahmin edicilerden bahsedilecektir.
- Artık günümüzde simülasyon diline entegre olan bir çok yazılım bu tahmin edicileri hesaplamakta kullanılmaktadır.

ÖN İSTATİSTİK: ÖRNEKLEM ORTALAMASI VE ÖRNEKLEM VARYANSI

- Bir çok durumda örneklem ortalaması yada örneklem ortalaması ve örneklem varyansı sınanan bir dağılımın parametrelerini tahmin etmede kullanılmaktadır:
- Aşağıdaki örneklem ortalamasını ve varyansını hesaplamak üzere 3 denklem seti verilmiştir.

Eğer $X_1, X_2, \ldots, X_n$ gözlem değerleri n boyutlu bir örneklem ise örneklem ortalaması $(\overline{X})$ aşağıdaki gibi hesaplanır:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

ve örneklem varyansı S^2 aşağıdaki gibidir:

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

Eğer veri kesikli ve frekans dağılımı içerisine gruplandırılmışsa yukarıdaki denklemler hesaplama kolaylığı sağlamak için modifiye edilebilir. Örneklem ortalaması

$$\overline{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n}$$

Ve örneklem varyansı

$$S^2 = \frac{\sum_{j=1}^{k} f_j X_j^2 - n\overline{X}^2}{n-1}$$

Burada k farklı X değerlerinin sayısını ve f_j ise X'in gözlenen X_j değerlerinin sayısını vermektedir.

ÖRNEK(GRUPLANDIRILMIŞ VERİ)

Her periyottaki varışlar	*Frekans*	*Her periyottaki varışlar*	*Frekans*
0	12	6	7
1	10	7	5
2	19	8	5
3	17	9	3
4	10	10	3
5	8	11	1

Yukarıdaki tablodaki veriler analiz edilirse $n = 100$, $f_1 = 12$, $X_1 = 0$, $f_2 = 10$, $X_2 = 1$,..., $\sum_{j=1}^{k} f_j X_j = 364$ ve $\sum_{j=1}^{k} f_j X_j^2 = 2080$ bulunur. O halde

$$\overline{X} = \frac{364}{100} = 3{,}64$$

ve

$$S^2 = \frac{2080 - 100(3{,}64)^2}{99} = 7{,}63$$

Örneklem standart sapması $S = \sqrt{7{,}63} = 2{,}76$.

TAHMİN EDİCİLER

Dağılım	*Parametre(ler)*	*Önerilen Tahmin Edici(ler)*
Poisson	α	$\hat{\alpha} = \overline{X}$
Üstel	λ	$\hat{\lambda} = \frac{1}{\overline{X}}$
Gamma	β, θ	$\hat{\beta}$ $\hat{\theta} = \frac{1}{\overline{X}}$
Normal	μ, σ^2	$\hat{\mu} = \overline{X}$ $\hat{\sigma}^2 = S^2$ $(yansız)$
Lognormal	μ, σ^2	$\hat{\mu} = \overline{X}$ (verinin ln'ini aldıktan sonra) $\hat{\sigma}^2 = S^2$ (verinin ln'ini aldıktan sonra)
Weibul $v = 0$	α, β	$\hat{\beta}_0 = \frac{\overline{X}}{S}$ $\hat{\beta}_j = \hat{\beta}_{j-1} - \frac{f(\hat{\beta}_{j-1})}{f'(\hat{\beta}_{j-1})}$ $\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\hat{\beta}}\right)^{1/\hat{\beta}}$

ÖRNEK(POISSON DAĞILIMI)

Her periyottaki varışlar	*Frekans*	*Her periyottaki varışlar*	*Frekans*
0	12	6	7
1	10	7	5
2	19	8	5
3	17	9	3
4	10	10	3
5	8	11	1

- Yukarıdaki verileri dikkate alalım. Verilerin histogramı çizildiğinde parametresi bilinmeyen α olan Poisson dağılımından şüphelenilmektedir. α parametreli Poisson dağılımının tahmin edicisinin $\hat{\alpha} = \overline{X}$ olduğunu biliyoruz. Böylece $\hat{\alpha} = 3{,}64$ olur (daha önceki örnekte bulunmuştu). Gerçek Poisson dağılımının ortalamasının ve varyansının birbirine eşit olduğunu hatırlayalım. Örneklem standart sapmasının $S^2 = 7{,}63$ olarak bulunmuştu. Ancak örneklem ortalaması ve varyansı rassal değişkenler olduğu için bunların tam olarak birbirine eşit çıkması asla beklenmemelidir.

ÖRNEK (Lognormal Dağılım)

- Bir portföydeki 10 yatırımın geri dönüş oranları yüzde olarak 18.8, 27.9, 21.0, 6.1, 37.4, 5.0, 22.9, 1.0, 3.1 ve 8.3'tür. Bu verinin lognormal modelinin parametrelerini tahmin etmek için ilk olarak verilerin doğal logaritmalarını alınmalıdır. O halde logaritma alındıktan sonra veriler 2.9, 3.3, 3.0, 1.8, 3.6, 1.6, 3.1, 0, 1.1 ve 2.1. Öyleyse $\hat{\mu} = \overline{X} = 2.3$ ve $\hat{\sigma}^2 = S^2 = 1.3$ bulunur.

ÖRNEK (Normal Dağılım)

- Normal dağılımın parametreleri olan μ ve σ^2, $\overline{X}$ ve S^2 tahmin edicilerinden bulunabilir. Kapıların montaj zamanlarına ilişkin q-q grafiğinin çizildiği örnekte histogram montaj zamanlarının normal dağılıma benzediğini göstermektedir. O halde $\hat{\mu} = \overline{X} = 99.9865$ ve $\hat{\sigma}^2 = S^2 = (0.2832)^2\ sn^2$ olur

UYUM İYİLİĞİ TESTİ

- Uyum iyiliği testleri potansiyel bir girdi modelinin uygunluğunun değerlendirilmesinde faydalı bir rehberdir.
- Ancak gerçek uygulamalarda doğru tek bir dağılım olmadığı için bu testlerinin sonuçlarının kölesi de olmamak gerekir.
- Örneklem hacminin büyüklüğünün etkisini anlamak özellikle önemlidir.
- Eğer çok az veri mevcutsa uyum iyiliği testi herhangi bir aday dağılımın muhtemelen reddetmeyecektir.
- Eğer veri sayısı çoksa uyum iyiliği testleri muhtemelen tüm aday çözümleri reddedecektir.

Kİ-KARE TESTİ

- Bu test verinin histogramı ile aday yoğunluk veya kütle fonksiyonunun şeklinin karşılaştırılması prensibiyle çalışmaktadır. Bu test parametreler maksimum olasılık fonksiyonu ile tahmin edildiğinde büyük örneklem hacimleri için sürekli ve kesikli dağılımlar için kullanımı uygundur. Test prosedürü n tane veriyi k tane sınıf içerisine sokarak başlamaktadır. Test istatistiği

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Burada O_i, i'nci sınıf aralığındaki gözlenen frekans değeri ve E_i ise o sınıf aralığındaki beklenen frekans değeridir. Her bir sınıf aralığı için beklenen frekans $E_i = np_i$ şeklinde hesaplanmaktadır. Burada p_i i'nci sınıf aralığına ilişkin teorik olasılık değeridir.

s örneklem istatistiğiyle tahmin edilen teorik dağılımın parametre sayısını göstermek üzere χ_0^2'nin serbestlik derecesi $k-s-1$ olan Ki-Kare dağılımına uyduğu gösterilebilir. Hipotezler aşağıdaki gibi kurulur:

- H_0: X rassal değişkeni parametresi(parametreleri) tahmin edici(tahmin ediler) ile hesaplanmış teorik dağılıma uymaktadır.
- H_1: X rassal değişkeni teorik dağılıma uymamaktadır.

Kritik değer $\chi^2_{\alpha,k-s-1}$ Ki-Kare tablosundan bulunabilir. Eğer $\chi_0^2 > \chi^2_{\alpha,k-s-1}$ ise sıfır hipotezi H_0 reddedilir.

E_i'nin minimum değerine ilişkin genel bir kabul olmamasına karşın 3, 4 ve 5 değerleri yaygın olarak kullanılmaktadır. Eğer E_i değeri çok küçükse komşu sınıf aralıklarının beklenen frekanslarıyla birleştirilebilir. İlgili O_i değerleri de birleştirilmelidir ve k birleştirilen her bir sınıf için 1 azaltılmalıdır.

- Eğer test edilen dağılım kesikli ise rassal değişkenin her bir değeri, minimum-beklenen hücre frekansı gereksinimini karşılamak için komşu sınıf aralıklarıyla birleştirilmeyecekse bir sınıfı temsil etmelidir. Kesikli durum için komşu sınıfların birleştirilmesi gerekmiyorsa

$$p_i = p(x_i) = P(X = x_i)$$

- Diğer durumda p_i uygun komşu sınıfların olasılıkları toplanarak bulunur.
- Eğer test edilen dağılım sürekli ise, a_{i-i} ve a_i i'nci sınıf aralığının uç noktalarını göstermek üzere sınıf aralıkları $[a_{i-1}, a_i)$ şeklinde verilir. Varsayılan yoğunluk fonksiyonunun $f(x)$ veya birikimli dağılım fonksiyonunun $F(x)$ olduğu sürekli durum için p_i aşağıdaki gibi hesaplanır:

$$p_i = \int_{a_{i-1}}^{a_i} f(x)dx = F(a_i) - F(a_{i-1})$$

- Kesikli durum için sınıf aralıklarının sayısı komşu sınıfların birleştirilmesinden sonra (eğer gerekliyse) elde edilen sınıfların sayısı ile belirlenir.

- Ancak sürekli durum için sınıf sayısı açıkça belirtilmelidir. Takip edilecek genel bir kural olmamasına karşın sürekli veri için sınıf aralıklarının sayısının belirlenmesinde aşağıdaki tablodan yardım alınabilir.

Örneklem Büyüklüğü, n	*Sınıf Aralıklarının Sayısı, k*
20	Ki-Kare testini kullanma
50	5-10
100	10-20
>100	$\sqrt{n} - n/5$

Ki-Kare testi şu adımlar gerçekleştirilerek yapılır.

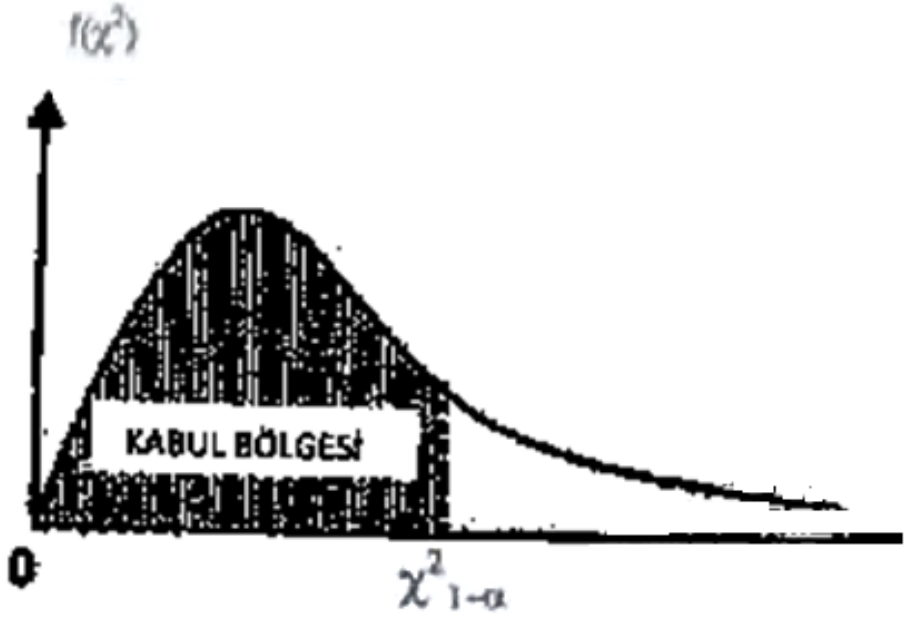
1. Örnek verilerin tüm parametrelere verilmiş veya tahmin edilmiş olan bir dağılımdan geldiği hipotezi kurulur.
2. Test edilen dağılımın aralığı, m adet alt bölgeye ve alt aralığa bölünür.
3. Her bir aralığa denk gelen teorik frekans değeri belirlenir. Burada her bir gruba düşen frekans sayısının veya örnek adedinin 5'ten az olmaması sağlanır.
4. Qj örnekleme frekansı, Ej dağılımın teorik frekansı olmak üzere Ki-Kare değeri aşağıdaki bağıntı ile hesaplanır.

$$\text{Ki} - \text{Kare} = \sum_{j=1}^{m} \frac{(Q_j - E_j)^2}{E_j}$$

5. Belirlenen güvenirlilik değerine ve serbestlik derecesine göre tablodan Ki-kare değeri okunur.

   Serbestlik Derecesi = n-(test edilen dağılımın parametre sayısı)-1

6. Şayet hesaplanan Ki-kare değeri, tablodan okunan değerden küçükse hipotez KABUL aksi halde hipotez RED edilir.



ÖRNEK (POISSON VARSAYIMINA UYGULANAN KI-KARE TESTİ)

- Aşağıdaki tablodaki araçların varışlarına ilişkin verileri yeniden hatırlayalım. Verilere ilişkin histogram verilerin poisson dağılımından gelebileceğini göstermişti ve $\hat{\alpha} = 3{,}64$ olarak bulunmuştu. Böylece aşağıdaki hipotezler geliştirilebilir.

Her periyottaki varışlar	*Frekans*	*Her periyottaki varışlar*	*Frekans*
0	12	6	7
1	10	7	5
2	19	8	5
3	17	9	3
4	10	10	3
5	8	11	1

- H_0: rassal değişken Poisson dağılımına uymaktadır.
- H_1: rassal değişken Poisson dağılımına uymamaktadır.

Poisson dağılımının olasılık kütle fonksiyonunun aşağıdaki gibi olduğunu hatırlarsak

$$p(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^{x}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & diğer\ durumda \end{cases}$$

$\alpha = 3{,}64$ için farklı x değerlerine ilişkin olasılıklar yukarıdaki denklem kullanılarak aşağıdaki gibi bulunabilir.

$p(0) = 0{,}026$
$p(1) = 0{,}096$
$p(2) = 0{,}174$
$p(3) = 0{,}211$
$p(4) = 0{,}192$
$p(5) = 0{,}140$
$p(6) = 0{,}085$
$p(7) = 0{,}044$
$p(8) = 0{,}020$
$p(9) = 0{,}008$
$p(10) = 0{,}003$
$p(11) = 0{,}001$

x_i	Gözlenen Frekans, O_i		Beklenen Frekans, E_i		$\frac{(O_i - E_i)^2}{E_i}$
0	12	22	2,6	12,2	7,87
1	10		9,6		
2	19		17,4		0,15
3	17		21,1		0,80
4	10		19,2		4,41
5	8		14,0		2,57
6	7		8,5		0,26
7	5	17	4,4	7,6	11,62
8	5		2,0		
9	3		0,8		
10	3		0,3		
11	1		0,1		
	100		100,0		27,68

- $\chi_0^2 = 27{,}68$ olarak bulunmuştur. Tablo değeri için serbestlik derecesi $k - s - 1 = 7 - 1 - 1 = 5$ bulunur. Veriden tahmin edilen tek bir parametre olduğu için $(\hat{\alpha})$ burada $s = 1$ alınmıştır. 0,05 önem seviyesinde kritik değer $\chi^2_{0.05,5} = 11{,}1$ olduğu görülür. Böylece 0.05 anlam seviyesinde H_0 hipotezi reddedilir.

ÖRNEK (ÜSTEL DAĞILIM İÇİN Kİ-KARE TESTİ)

Daha önce arıza verileri analiz edilmişti ve verilerin histogramının üstel dağılıma benzediği görülmüştü. Parametre tahmin edicisinin $\hat{\lambda} = 1/\overline{X} = 0{,}084$ olarak bulunmuştur. O halde aşağıdaki hipotezler kurulabilir:

- H_0: rassal değişken üstel dağılıma uymaktadır.
- H_1: rassal değişken üstel dağılıma uymamaktadır.

Eşit olasılıklı aralıklara Ki-Kare testi uygulamak için sınıf aralıklarının uç noktalarının belirlenmesi gerekmektedir. Sınıf sayısının $n/5$'ten küçük veya eşit olmalıdır. Burada $n = 50$ olduğundan $k \leq 10$ olmalıdır.

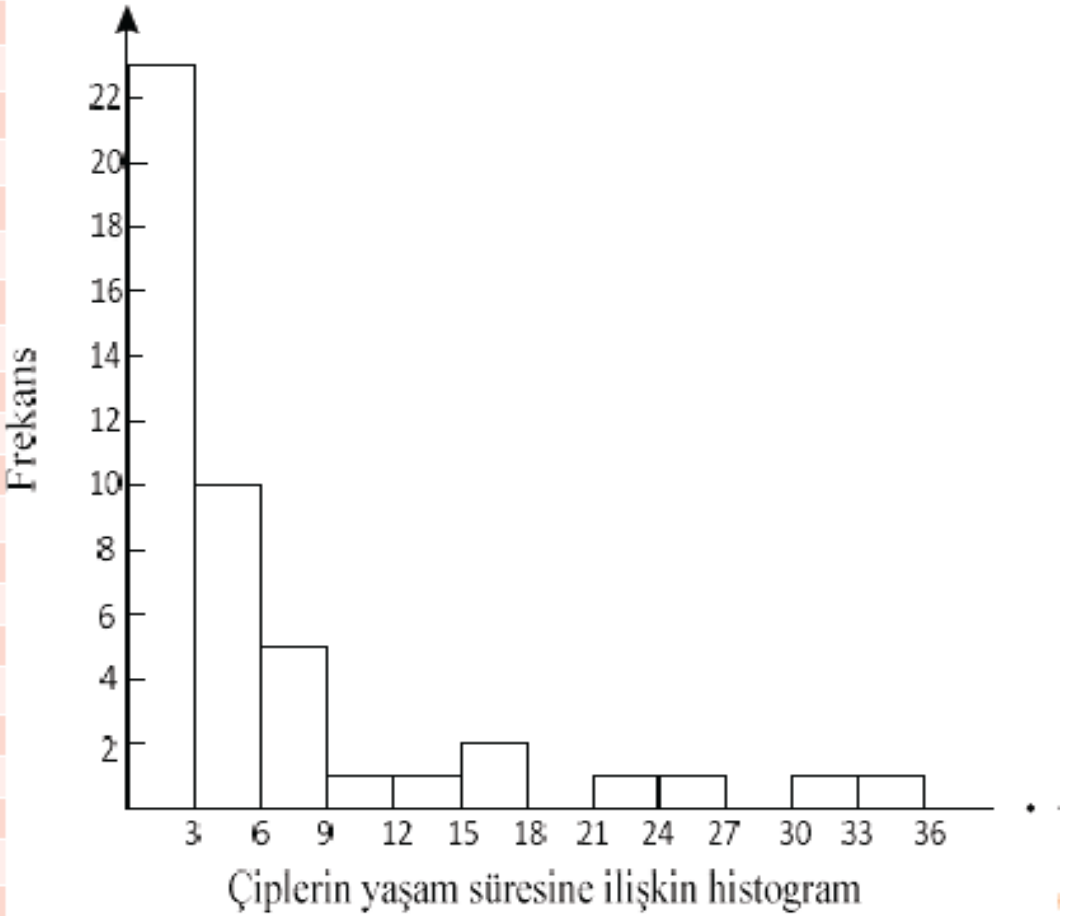
Tablodan sınıf sayısının 5 ile 10 arasında olması gerektiği görülmektedir. Mesela $k = 8$ olsun böylece her bir aralık $p = 0{,}125$ olasılık değerine sahip olacaktır. Her bir aralığın uç noktaları üstel dağılımın bdf'sinden aşağıdaki gibi hesaplanabilir.

$$F(a_i) = 1 - e^{-\lambda a_i}$$

Normal voltajın 1.5 katı verilerek elektronik çiplerden oluşan bir rassal örnekleme ömür testleri yapılmış ve yaşam süreleri (ya da bozulana kadar geçen süreler) gün cinsinden kaydedilmiştir:

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562

Çip Ömrü (Gün)	Frekans
$0 \leq x_j < 3$	23
$3 \leq x_j < 6$	10
$6 \leq x_j < 9$	5
$9 \leq x_j < 12$	1
$12 \leq x_j < 15$	1
$15 \leq x_j < 18$	2
$18 \leq x_j < 21$	0
$21 \leq x_j < 24$	1
$24 \leq x_j < 27$	1
$27 \leq x_j < 30$	0
$30 \leq x_j < 33$	1
$33 \leq x_j < 36$	1
.	.
.	.
.	.
$42 \leq x_j < 45$	1
.	.
.	.
.	.
$57 \leq x_j < 60$	1
.	.
.	.
.	.
$78 \leq x_j < 81$	1
.	.
.	.
.	.
$144 \leq x_j < 147$	1



Çiplerin yaşam süresine ilişkin histogram

Burada a_i değeri i. aralık için $(i = 1,2,\dots,k)$ uç nokta değeridir. $F(a_i)$ değeri o'dan a_i'ye kadar birikimli alan olduğundan, $F(a_i) = ip$ yazılabilir. Öyleyse

$$ip = 1 - e^{-\lambda a_i}$$

veya

$$e^{-\lambda a_i} = 1 - ip$$

Her iki tarafın logaritması alınıp a_i değeri çekildikten sonra üstel dağılım için k tane eşit olasılıklı aralığın uç noktaları aşağıdaki gibi bulunur:

$$a_i = -\frac{1}{\lambda} ln(1 - ip), \qquad i = 0,1,\dots,k$$

λ'nın değerini önemsemeksizin yukarıdaki denklemde $a_0 = 0$ ve $a_k = \infty$ olacaktır. $\hat{\lambda} = 0,084$ ve $k = 8$ için a_1 değeri yukarındaki eşitlikten bulunabilir:

$$a_1 = -\frac{1}{0,084} ln(1 - 0,125) = 1,590$$

- Tablodan görüleceği gibi $\chi_0^2 = 39{,}6$ bulunmuştur. Serbestlik derecesi $k - s - 1 = 8 - 1 - 1 = 6$ 'dır. $\alpha = 0{,}05$ için kritik tablo değeri $\chi^2_{0.05,6} = 12{,}6$'dır. $\chi_0^2 > \chi^2_{0.05,6}$ olduğundan sıfır hipotezi reddedilir. Ayrıca $\alpha = 0{,}01$ için $\chi^2_{0.01,6} = 16{,}8$ olduğu için sıfır hipotezi 0,01 önem seviyesinde de reddedilecektir.

Sınıf Aralığı	*Gözlenen Frekans,* O_i	*Beklenen Frekans,* E_i	$\frac{(O_i - E_i)^2}{E_i}$
[0, 1,590)	19	6,25	26,01
[1,590, 3,425)	10	6,25	2,25
[3,425, 3,425)	3	6,25	0,81
[5,595, 3,425)	6	6,25	0,01
[8,252, 3,425)	1	6,25	4,41
[11,677, 3,425)	1	6,25	4,41
[16,503, 3,425)	4	6,25	0,81
[24,755, ∞)	6	6,25	0,01
	50	50	39,6

KOLMOGOROV-SMIRNOV UYUM İYİLİĞİ TESTİ

- Kolmogorov-Smirnov Testi özellikle örneklem hacmi küçükken ve veriden bir parametre tahmini yapılmadığı durumda etkin olarak kullanılır.
- Kolmogorov-Smirnov testi verilerin üstel dağılıma uyup uymadığı test ederken herhangi bir özel tablo kullanmaz.

KOLMOGOROV-SMIRNOV testi şu adımlar gerçekleştirilerek yapılır.

1 Örnek verilerin tüm parametrelere verilmiş veya tahmin edilmiş olan bir dağılımdan geldiği hipotezi kurulur.

2. Test edilen dağılımın aralığı, m adet alt gruba veya alt aralığa bölünür.

3 Test edilen dağılımın gruplara ait teorik birikimli olasılık değerleri bulunur. (F(y))

4. Örneklemden elde edilen dağılımın birikimli olasılık değerleri bulunur. ($F_0(y)$)

5. Her alt aralık için $|F(y) - F_0(y)|$ değerlerini belirlenir.

6. D_N değeri belirlenir.

$$D_N = \underset{j=1...m}{enb}\left\{\left|F(y) - F_0(y)\right|\right\}$$

7 D_{Nt} belirlenir. Bu değer tekli örnekleme yapılıyor ise

% 90 güvenirlilik için $D_{Nt} = \frac{1.22}{\sqrt{n}}$ bağıntısı ile,

% 95 güvenirlilik için $D_{Nt} = \frac{1.36}{\sqrt{n}}$ bağıntısı ile,

% 99 güvenirlilik için $D_{Nt} = \frac{1.63}{\sqrt{n}}$ bağıntısı ile hesaplanabilir.

8. Şayet belirlenen D_N değeri, D_{Nt} değerden küçükse hipotez KABUL aksi halde hipotez RED edilir.

ÖRNEK (ÜSTEL DAĞILIM İÇİN KOLMOGOROV-SMIRNOV TESTİ)

Varsayalım ki 100-dakikalık bir zaman dilimi içerisinde 50 adet varışlar arası zaman (dk cinsinden) oluşum sırasına göre toplanmış olsun:

0,44	0,53	2,04	2,74	2,00	0,30	2,54	0,52	2,02	1,89
1,53	0,21	2,80	0,04	1,35	8,32	2,34	1,95	0,10	1,42
0,46	0,07	1,09	0,76	5,55	3,93	1,07	2,26	2,88	0,67
1,12	0,26	4,57	5,37	0,12	3,19	1,63	1,46	1,08	2,06
0,85	0,83	2,44	1,02	2,24	2,11	3,15	2,90	6,58	0,64

Sıfır hipotezi ve alternatif hipotezler aşağıdaki gibi kurulur:

- H_0: Varışlar arası zaman üstel dağılıma uymaktadır.
- H_1: Varışlar arası zaman üstel dağılıma uymamaktadır.
- Veriler 0 ile $T = 100$ dakika aralığında toplanmıştır.

- Varsayılan dağılımın varışlar arası zamanı $\{T_1, T_2, \dots\}$ üstel ise varış zamanlarının (o, T) aralığında düzgün dağıldığı gösterilebilir.
- Varış zamanları $T_1, T_1 + T_2, T_1 + T_2 + T_3, \dots T_1 + \cdots + T_{50}$ varışlar arası zamanlar eklenerek elde edilebilir.
- Daha sonra varış zamanları $(0, 1)$ aralığına normalize edilir ve Kolmogorov-Smirnov testi uygulanabilir.
- $(0, 1)$ aralığında noktalar şu şekilde olur: $[T_1/T,\ (T_1 + T_2)/T, \dots, (T_1 + T_2 + T_3)/T]$

0,0044	0,0097	0,0301	0,0575	0,0775	0,0805	0,1059	0,1111	0,1313	0,1502
0,1655	0,1676	0,1956	0,1960	0,2095	0,2927	0,3161	0,3356	0,3366	0,3508
0,3553	0,3561	0,3670	0,3746	0,4300	0,4694	0,4796	0,5027	0,5315	0,5382
0,5494	0,5520	0,5977	0,6514	0,6526	0,6845	0,7008	0,7154	0,7262	0,7468
0,7553	0,7636	0,7880	0,7982	0,8206	0,8417	0,8732	0,9022	0,9680	0,9744

- Rassal sayıların düzgünlük testlerine ilişkin yolu takip ettiğimizde $D^+ = 0{,}1054$ ve $D^- = 0{,}0080$ bulunur.
- Böylece Kolmogorov-Smirnov istatistiği $D = max(0{,}1054, 0{,}0080) = 0{,}1054$ bulunur. $\alpha = 0{,}05$ ve $n = 50$ için tablodan kritik D değeri $D_{0{,}05} = 1{,}36/\sqrt{n} = 0{,}1923$ bulunur.
- $D = 0{,}1054$ olduğundan varışlar arası zamanın üstel dağılımından geldiği hipotezi reddedilemez.