CHAPTER THREE

CHOOSING PROBABILITY DISTRIBUTIONS TO MODEL REAL WORLD PHENOMENA

3.1 Common Probability Distributions

Before we can perform simulations of real world phenomena, such as customers entering a bank and waiting on a line (queue), we need to select appropriate probability distributions that match the real world data.

Let us examine several discrete probability distributions and then several continuous probability distributions. Commonly occurring discrete distributions are the binomial and the Poisson.

Binomial Distribution

As previously noted, if p = probability of an event and 1-p = probability that the event will not occur, and x = the number of times the event will occur in n trials, then

$$f(x) = P(X\text{-}x) = \binom{n}{x} p^x (1 - p)^{n-x},\ x = 0, 1, \ldots, n.$$

The mean of the binomial distribution is $\mu = np$, the variance $= \sigma^2 = np(1\text{-}p)$. These statistics enable you to estimate using the z table the probability associated with an outcome occurring - for example, the probability of between 50 and 80 heads occurring in 100 trials.

This falls within the domain of the binomial distribution, since P = 1/2 for each head and each flip is independent of the previous toss.

μ = mean = np = 100 (1/2) = 50

σ^2 = variance = np (1-p) = 25

$\sigma = 5 =$ standard deviation

To be precise, 49.5 rounds to 50 and 80.4 rounds to 80. Therefore, we will find the probability:

$$P(49.5 < x < 80.5)^*$$

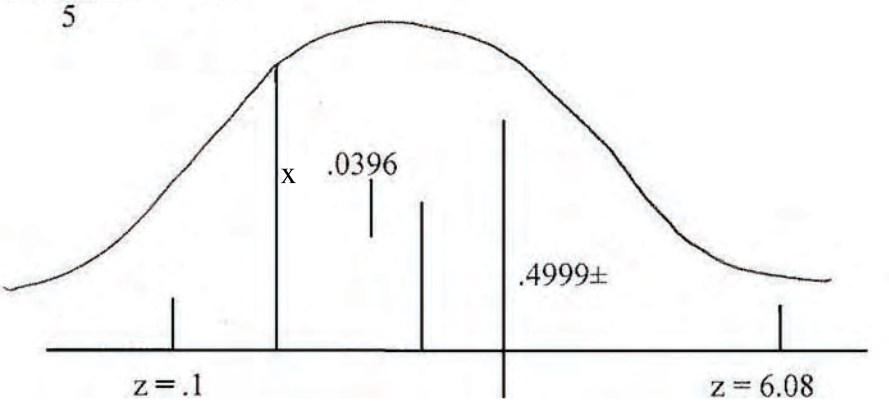
Change to z scores: for $x = 49.5$

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{49.5 - 50}{5} = \frac{-.5}{5} = -.1$$

for $x = 80.4$

$$z = \frac{80.4 - 50}{5} = 6.08$$



*This adjustment is necessary to apply the z table, which is the probability of a continuous random variable, to the binomial distribution (a discrete random variable's probability distribution).

Use the z table to obtain the probabilities:

$P(-.1 < z < 0) = .0398$

$P(0 < z < 6.08) = .4999+$

$P(-.1 < z < 6.08) = .0398 + .4999+ = .5398$

Therefore, there is a 53.98% chance of obtaining between 50 and 80 heads in 100 coin flips. The slight discrepancy between 50 and 49.5 illustrates the difficulty of approximating discrete phenomena (# of heads) with the area under the normal curve between two points (continuous random variables).

We clearly cannot obtain 49.5 heads in 100 coin flips. However, for large n, the discrepancy becomes minor. Another approach to use if n is at least 20 and p is at most .05 is the ***Poisson distribution***. In fact, as $n \to \infty$ and $p \to 0$, the binomial distribution approaches the Poisson distribution as a limit. We will discuss the Poisson distribution in greater detail later on.

Normal Distribution

We have frequently mentioned the ***normal distribution***. Its density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

where μ = mean and σ = standard deviation.

The distribution function, which requires numerical analysis to calculate, is defined as:

$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} \, dt$$

When one looks up values such as our earlier value of $P(-.1 < z < 6.08)$, we are relying on the table that has already computed:

$$\int_{-.1}^{0} e^{-(x-\mu)^2/2\sigma^2}\,dx + \int_{0}^{6.08} e^{-(x-\mu)^2/2\sigma^2}\,dx$$

$$= .0398 + .4999+ = .5398$$

To illustrate one of the basic formulas of numerical analysis applied to the integral of a simplified normal distribution, let $\mu = 0,\ \sigma = 1$. We have $\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx$, which cannot be computed by elementary methods. One of the common formulas for such problems is Simpson's rule which is written as follows:

$$\int_{x_0}^{x_2} f(x)\,dx = h/3\ [f(x^0) + 4\,f(x_1) + f(x_2)]$$

with Error $\leq h^5/90\ f^{(4)}\ (E),\ x_0 < E < x_2$.

For our problem $x_0 = 0, x_1 = 1,\ h = \frac{b-a}{n} = \frac{x_2 - x_0}{2} = 1/2$

and $f(x) = e^{-x^2/2}$. Therefore,

$$\int_0^1 \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}}\,\frac{1/2}{3}\,[e^{-0^2/2} + 4e^{-(1/2)^2/2} + e^{-1^2/2}]$$

$$\int \approx 1/6\ [1 + 4e^{-1/8} + e^{-1/2}]$$

$$\int \approx 1/6\ [1 + 4(.8825) + (.6065)] = .8561$$

$$\frac{1}{\sqrt{2\pi}} \int \approx \frac{.8561}{2.5066} \approx .3415$$

If you look up the z value corresponding to z = 1.0, you find a probability measure of .3413. Our answer is accurate to 10^{-3}. The error analysis, which is an essential part of numerical analysis, requires that you compute the fourth derivative of f(x). You then substitute some E between x_0 and x_2 to overstate or maximize the error. This way you are safe to say that the

actual error is less than your estimate. To complete the error analysis for this problem, the four derivatives of f are computed as follows:

$$f(x) = e^{-x^2/2}$$

$$f' = (e^{-x^2/2})(-x)$$

$$f'' = (e^{-x^2/2})(-1) - x(e^{-x^2/2})(-x) = (x^2-1)e^{-x^2/2}$$

$$f''' = (x^2 - 1)(e^{-x^2/2})(-x) + (e^{-x^2/2})(2x) = (-x^3 + 3x)e^{-x^2/2}$$

$$f^4 = (-x^3 + 3x)(e^{-x^2/2})(-x) + (e^{-x^2/2})(-3x^2 + 3) = (x^4 + 3)e^{-x^2/2}$$

To return to the error formula, $E \leq h^5/90\, f^{(4)}(E)$. Let $h = 1/2$. For $(x^4 + 3)$, let $x = 0$ to overstate error. Similarly, for $e^{-x^2/2}$, let $x = 0$. We have to substitute a number between 0 and 1 (x_0 and x_2) for E, but it need not be the same value in each parentheses of a complicated product or quotient. But this art is an important skill that should be developed in an entire course devoted to Numerical Analysis. Our final error is computed as follows:

$$E \leq \frac{(1/2)^5}{90}(4)(e^{-0}) = .0010$$

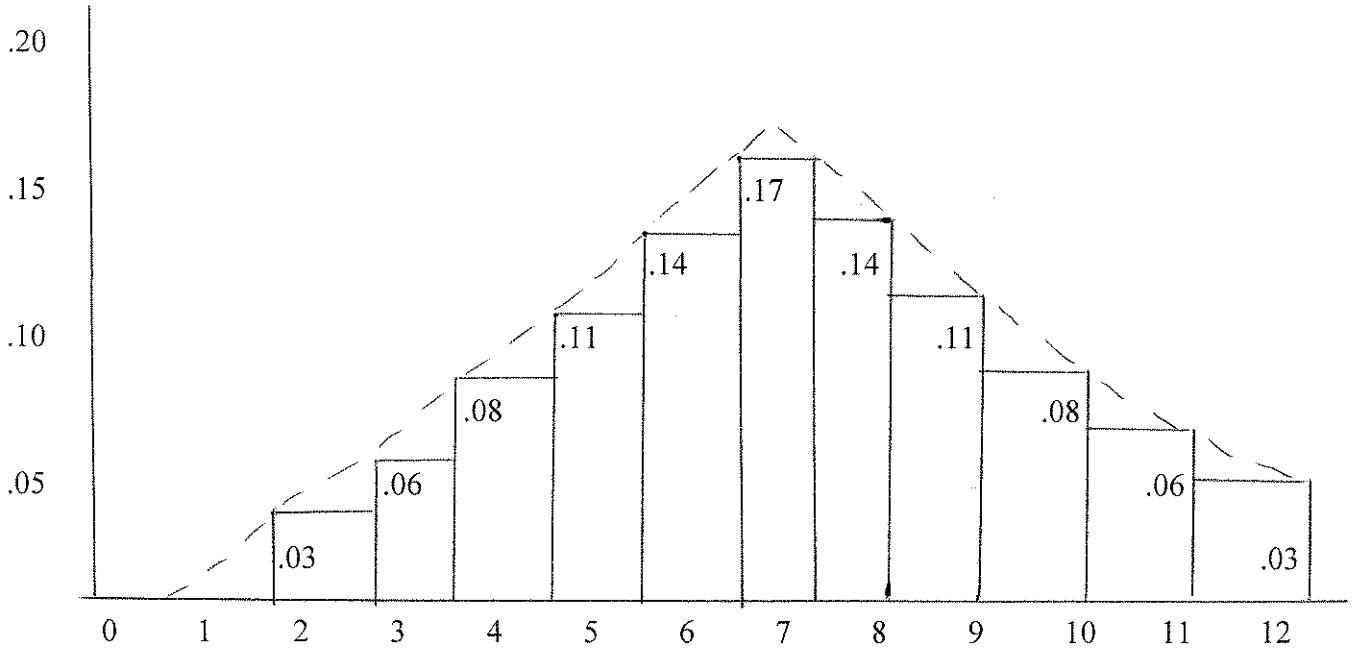
This error estimate is in harmony with our actual error which is .0002. We nearly always overstate error in numerical analysis. This is because many numerical procedures are estimating real world phenomena, and we want to be especially careful in our analyses which often relate to people's safety and well-being.

The reason that we were able to approximate our earlier probabilities was that the binomial distribution can be approximated by the normal distribution if n is large. That is, if n is sufficiently large and p and (1-p) are not too small, the binomial distribution can be approximated by the standardized normal random variables with $z = \frac{x - \mu}{\sigma}$ where $\mu = np$, $\sigma = \sqrt{np(1-p)}$. This is the formula that we used in our previous example.

To illustrate further the connection between the binomial distribution and the normal curve, consider a histogram which graphically represents the relative frequency of outcomes from a dice throw. We will postpone the way to construct histograms until Section 3.3. We know if we throw two dice the following probabilities are computable from simple analysis of the 36 possibilities in the sample space:

P(2)	=	1/36	P(7)	=	1/6
P(3)	=	1/18	P(8)	=	5/36
P(4)	=	1/12	P(9)	=	1/9
P(5)	=	1/9	P(10)	=	1/12
P(6)	=	5/36	P(11)	=	1/18
			P(12)	=	1/36

The histogram is drawn below:



The normal curve is added above by a dotted line approximation and offers geometrical insight into the connection between the normal curve and the relative frequence (or calculated probability) of the binomial event of a dice throw.

Poisson Distribution

Another commonly used distribution in mathematical modeling is the ***Poisson distribution***. It was discovered by S.D. Poisson almost two hundred years ago, and it is safe to say that he never anticipated that it could eventually be used to model interarrival time for drive-up banking depositors.

The Poisson distribution is discrete and is given by:

$$f(x) = P(X - x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots\ldots$$

λ is a positive real number

$e^{-\lambda}$ can be computed with most scientific calculators or looked up in tables.

The Poisson distribution has the extremely unusual property that both the mean μ and the variance σ^2 equal λ.

One application of the Poisson distribution is when you are dealing with rare events. For example, suppose $n > 50$ and $p < 1/10$. The Poisson distribution can give you a good estimate of the probability of rare occurrences. Let $p = .05$ be the probability of contracting a rare disease in a particular environment. Let $n = 200$. The population under consideration equals 200. The various values of x and corresponding probabilities represent x occurrences in the total population. To find the probability that exactly four people will contract the disease, use the Poisson distribution:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda = np = 200\,(.05) = 10$$

$$x = 4$$

$$f(4) = \frac{10^4 e^{-10}}{4!} = .019$$

We conclude that the probability of exactly four people contracting the disease is approximately .019. The Poisson distribution approximates the binomial distribution if n is large (e.g., if $n > 50$) and p is close to 0 (e.g., if $p \leq .05$). The approximation is excellent if $n \geq 100$ and $np \leq 10$.

Central Limit Theorem

One of the most remarkable results in mathematics is that a great many distributions approach the normal distribution and therefore can be approximated by a single curve.

The Central Limit Theorem states:

> If $x_1, x_2, \ldots x_n$ are independent random variables with the same probability function (discrete or continuous) having finite means μ and variance σ^2, and if $s = x_1 + x_2 + \ldots x_n$, then $\frac{s_n - n\mu}{\sigma/\sqrt{n}}$ is normal (as $n \to \infty$).

This theorem was critical in our earlier material in confidence intervals and hypothesis testing.

Uniform Distribution (Continuous)

When in doubt as to which distribution to use for a continuous random variable, consider the uniform distribution. Whether one wants a random number from 0 to 1 or to model a random time from 8 A.M. to 4 P.M. for customers to enter a bank, researchers frequently use the uniform distribution.

A continuous random variable Y is ***uniformly distributed*** on an interval $a \leq x \leq b$ if its density function is written:

$$f(x) = \frac{1}{b-a} \quad a \leq x \leq b$$

$$= 0 \qquad \text{elsewhere}$$

We have already shown (Section 2.5) that the mean of the uniform distribution is $\mu = E(x) = \frac{b+a}{2}$. This makes sense. Consider a group of 100 people who have heights uniformly distributed between 5'0" and 6'0". The mean would be approximately 5'6". In this example, a = 5'0", b = 6'0", and $\mu = E(x) = \frac{a+b}{2} = \frac{11'}{2} = 5.5' = 5'6"$. The variance of of the uniform distribution is given by $\sigma^2 = \frac{(b-a)^2}{12}$. We will leave its computation for an exercise.

Exponential Distribution (Continuous)

We have introduced this commonly used distribution in Section 2.2. The density function f(x) is defined by:

$$f(x) = \frac{1}{B} e^{-x/B} \qquad \text{if } x \geq 0$$

$$= 0 \qquad \text{elsewhere}$$

To compute the mean of the exponential distribution, we evaluate

$$E(x) = \int_0^\infty f(x)\ x\ dx$$

$$= \int_0^\infty x\ e^{-x/B}\ (1/B)\ dx$$

Use integration by parts: Let $u = x$

$$dv = \frac{1}{B}(e^{-x/B})$$

$$du = dx$$

$$v = -e^{-x/B}$$

$$\int_0^\infty x\, e^{-x/B}\left(\frac{1}{B}\right) dx = uv \Big|_0^\infty - \int_0^\infty v\, du$$

$$E(x) = -x\, e^{-x/B}\Big|_0^\infty - \int_0^\infty -e^{-x/B}\, dx$$

$\lim_{x\to\infty} -x\, e^{-x/B}$ is of the form $-\infty \cdot 0$.

Use L'Hopital's Rule:

$$\lim_{x\to\infty} \frac{-x}{e^{x/B}} = \lim_{x\to\infty} \frac{-1}{\frac{1}{B}e^{x/B}} = 0$$

Since $\lim_{x\to\infty} -x\, e^{-x/B} = 0$, we see that

$$E(x) = \int_0^\infty e^{-x/B}\, dx = -B\int_0^\infty e^{-x/B}\left(\frac{-1}{B}\right) dx$$

$$= -B\, e^{-x/B}\Big|_0^\infty = 0 - (-B) = B, \text{ so that}$$

$\mu = E(x) = B$ for the exponential distribution. We leave the calculation of the variance ($\sigma^2 = B^2$) as an exercise.

Geometric (Discrete)

We sometimes slightly modify the binomial density so that we are concerned with the number of binomial trials preceding (and including) the trial in which the first success occurs.

For example, what is the probability of selecting the first ace on the third trial (with replacement in a standard 52 card playing deck)? The probability of the event is simply the

probability of a non-ace on the first two trials and an ace on the third trial. P(an ace being chosen for the first time on the third trial) = 48/52 • 48/52 • 4/52. Let p = the probability of the ace = 1/13; let q = 1-p = the probability of the non-ace. In general for the geometric density,

$$f(x) = p\,q^{x-1}, \qquad x = 1, 2, \ldots$$

The mean of the geometric density is given by $\mu = 1/p$, the variance σ^2 equals q/p^2. Let us compute the mean of the geometric density.

Mean Geometric Density

$$E(x) = \mu = \sum_{j=1}^{\infty} j\, pq^{j-1}, \quad (q = 1-p)$$

p is a constant and can be factored.

$$E(x) = \mu = p \sum_{j=1}^{\infty} j\,(1-p)^{j-1}, \quad p < 1.$$

E(x) has the following infinite series representation:

$$E(x) = p\,[1\,(1-p)^0 + 2\,(1-p)^1 + 3\,(1-p)^2 + \ldots]$$

This is a difficult series to sum without the observation that

$$\frac{d}{dp}(1-p)^j = (-1)\,j\,(1-p)^{j-1}$$

This allows the substitution:

$$E(x) = (-1)\,p \sum_{j=1}^{\infty} \frac{d}{dp}(1-p)^j$$

We can differentiate the power series term by term and obtain the following:

$$E(x) = -p \frac{d}{dp} \sum_{j=1}^{\infty} (1-p)^j$$

We now use the formula for the sum of a geometric series with $r = (1-p) < 1$

$$s_x = \frac{a}{1-r} = \frac{1 - p}{1 - (1-p)} = \frac{1 - p}{p}$$

To complete the problem,

$$E(x) = (-p) \frac{d}{dp} \frac{(1 - p)}{p} = (-p) \frac{(-1)}{p^2} = 1/p$$

This completes the demonstration that $E(x) = 1/p$ for the geometric density. To illustrate its value, consider rolling two dice. Suppose we define success as rolling a seven. P(seven) = 1/6. The mean of the geometric density (if $p = 1/6$) for this problem equals

$\frac{1}{1/6} = 6$. This makes sense since six trials is a reasonable mean for the event of throwing the first six.

The variance for this example

$$\sigma^2 = q/p^2 = \frac{1 - 1/6}{(1/6)^2} = \frac{5/6}{1/36}$$

$$\sigma^2 = 30$$

$$\therefore \sigma = \sqrt{30}.$$

EXERCISES 3.1

1. Estimate using the normal density the probability of winning over 60 of 100 hands in blackjack if you play very well and win with 48% probability.

2. Show that for the binomial density $\sum f(x) = 1$.

3. Are the probabilities associated with the normal density uniform? Explain.

   [Does $P(1 < z < 2) = P(3 < z < h4)$?]

4. If the probability of being infected with a certain disease is $p = .01$, find the probability that of 250 people, exactly three have been infected. [Use the Poisson distribution.]

5. Show that the mean and variance of the Poisson distribution are both λ.

6. Why is independence critical to the central limit theorem? Consult a standard text on calculus-based probability to examine the proof of the CLT.

7. Show that the variance of the uniform distribution is $\sigma^2 = \frac{(b - a)^2}{12}$

8. Show that the variance of the exponential distribution is given by $\sigma^2 = B^2$.

   Hint: You must integrate by parts twice.

9. Show that the geometric density has variance $\sigma^2 = q/p^2$.

3.2 Less Frequently Used Probability Distribiutions

Discrete Distributions:

Multinomial Distribution

Suppose we have k mutually exclusive events $E_1, E_2, \ldots, E_x$ where the probabilities associated with the events are $p_1, p_2, \ldots, p_k$. Further we require $p_1 + p_2 + \ldots + p_k = 1$. If $x_1, x_2, \ldots, x_k$ are random variables defining the number of times that $E_1, E_2, \ldots, E_k$ will occur in n trials, we require $x_1, x_2, \ldots, x_k = n$. The multinomial distribution gives the joint probability that E_1 occurs n_1 times, E_2 occurs n_2 times... and E_k occurs n_k times. This joint probability function is written as follows:

$$P(x_1=n_1, x_2=n_2, \ldots x_k=n_k) = \frac{n!}{n_1!\ n_2!\ldots n_k!}\ p_1^{n_1} \cdot p_2^{n_2} \cdot \ldots\ p_k^{n_k}$$

The expected number of times $E_1, E_2, \ldots, E_k$ will occur in n trials is written as follows:

$$E(x_1) = n\ p_1, \quad E(x_2) = n\ p_2, \ldots, \quad E(x_k) = n\ p_k$$

Hypergeometric Distribution

Suppose we have a deck of playing cards and select cards without replacement. The probability of selecting an ace on the second card would change depending upon whether the first card was an ace or a non-ace.

If two events A and B are not independent, we have the probability of A and B different from the simple product of $P(A) \cdot P(B)$ which is $P(A \cap B)$ (or P(A and B)) for independent events. Independent events are events like coin flips where the outcome of one coin flip does not change the probability of the next. The concept of ***dependent*** events leads us to ***conditional probability***.

We define the conditional probability of B, given the occurrence of event A, as probability $P(B \mid A)$. The general probability formula for dependent events A and B is as follows:

$$P(A \cap B) = P(A) \cdot P(B \mid A).$$

For our example, let A = event of an ace on card 1. Let B = event of an ace on card 2. Note that event B is related to event A. If the first card were replaced and the deck then well shuffled, the events would be independent. But this is not the case. If the first card is an ace, it is removed from the deck and there are only three aces remaining in the deck. Therefore, $P(B \mid A) = 3/51$. The probability of two aces in a row would be

$$P(x=2) = 4/52 \cdot 3/51 = 1/13 \cdot 1/17 = 1/221$$

The hypergeometric distribution is the appropriate distribution to use if sampling is done without replacement and is written:

$$P(x=a) = \frac{\binom{b}{a}\binom{m}{n-a}}{\binom{b+m}{n}}$$

where a = number of successful trials, b = number of elements associated with a success [an ace is a success in our discussion], m = number of elements associated with a failure [n = 48 non-aces in our example], and n = number of trials.

For our example in which we calculated the probability of obtaining two aces, let us use the hypergeometric distribution but let n = 4 instead of 2. That is, calculate the probability of two aces in four trials (without replacement).

$$P(x=a) = P(x=2) = \frac{\binom{4}{2}\binom{48}{4-2}}{\binom{4+48}{4}}$$

$$P(x=2) = \frac{\binom{4}{2}\binom{48}{2}}{\binom{52}{4}}$$

$$P(x=2) = \frac{6 \cdot 1128}{270{,}725} = .025$$

This example of ***conditional probability*** and its relevance to the ***hypergeometric distribution*** highlights the issue of dependence in statistics. Virtually all results in statistics - including the hypothesis testing and confidence intervals reviewed in Chapter One - assumed independent random variables. As we develop more mathematical tools in this book, you will be introduced to a new area of research - dependent random variables. Since nearly all real-world research problems fail to meet rigid independence assumptions, modeling may eventually help us to adjust statistics to dependence. As undergraduates, you can become participants in the exciting world of mathematical research through the research problems presented in this book.

Continuous Distributions:

Gamma Distribution

If one were modeling repair time or time to service customers, the ***gamma distribution*** could be very useful.

Its density function is written:

$$f(x) = \frac{B^{-\alpha}\, x^{\alpha-1}\, e^{-x/B}}{\Gamma(\alpha)} \qquad x > 0 \text{ , where } \alpha > 0 \text{ and } B > 0$$

$$= 0 \qquad \text{elsewhere}$$

$\Gamma(\alpha)$ is called the ***gamma function***:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\, e^{-t}\, dt \quad \text{for } \alpha > 0.$$

Properties of the Gamma Function:

1. $\Gamma(k+1) = k!$ for $k = 0, 1, 2, \ldots$.

2. $\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$ for $\alpha > 0$.

3. $\Gamma(1/2) = \sqrt{\pi}$

4. For large values of n, $\Gamma(n+1) = \sqrt{2\pi n}\; n^n\; e^{-n}$

   [This is called Stirling's formula.]

   We could express this relationship between $\Gamma(n+1)$ and $\sqrt{2\pi n}\; n^n\; e^{-n}$ as follows:

$$\lim_{n\to\infty} \frac{\sqrt{2\pi n}\; n^n\; e^{-n}}{\Gamma(n+1)} = 1$$

It follows from (1) and (4) that a good approximation for n! for large n is given as follows:

$$n! = \sqrt{2\pi n}\; n^n\; e^{-n}$$

The mean of the gamma distribution is given by $\mu = \alpha B$; the variance $\sigma^2 = \alpha B^2$.

Normal Distribution

One of the most useful functions in statistics, by virtue of the central limit theorem, is the normal density. Its formula was given earlier as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\; e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

where $\sigma > 0$

We cannot easily compute

$$f(x) = \int_{-\infty}^{x} f(x)\; dx$$

and need numerical analysis to compute the distribution function. Fortunately this has been done and all statistics texts feature a table of probabilities associated with standard normal z scores.

The mean of the normal density is μ ; the variance is σ^2 .

Lognormal Density

A colleague who had been researching the stock market for several years told me that stock prices are lognormally distributed random variables. Of course, it took a lot of experience and goodness of fit tests, which is the next topic, to find this out.

Frequently it is helpful to draw a histogram using our data $x_1, x_2, \ldots, x_n$. We break up the range of the data into k disjoint equal intervals. We let our y_m value be the proportion of data values in the mth interval (x_m, x_{m+1}). The problem of selecting the number of intervals is tricky and a commonly used approach is ***Sturge's rule***. This rule says that the number of intervals x should be selected by the following formula:

$$k = 1 + \log_2 n$$

Round up. For example, if $k = 5.2$, then let $k = 6$.

According to Law and Kelton (1991), this rule may not be very useful and several different values of interval length should be experimented with until the histogram resembles some standard density function.*

Several graphs of useful continuous probability distributions are presented at the end of this section and include the uniform, exponential, gamma, normal and Weibull. If a density has a shape

---

*Law, A. and Kelton, W. *Simulation Modeling and Analysis*. McGraw Hill, New York, 1991, p. 361.

similar to the gamma but has a big "jump" next to 0, the lognormal is a good "guess." Its density is written:

$$f(x) = \frac{1}{x\sqrt{2\pi}\,\sigma} e^{-(\ell n\, x-\mu)^2/2\sigma^2} \quad \text{for } x > 0$$

$$= 0 \quad \text{elsewhere}$$

As in the case of the normal distribution, we cannot easily derive

$F(x) = \int_0^x f(x)\, dx$ and require numerical analysis for such calculations. The mean of the lognormal density is $e^{\mu + \sigma^2/2}$; the variance is $e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$.

Weibull Distribution

If one wished to model the time until a certain item failed to work, the Weibull would be a strong candidate. Its density is written:

$$f(x) = k\, x^{B-1} e^{-\alpha x^B} \quad \text{for } x > 0$$

$$= 0 \quad \text{elsewhere where } 2 > 0 \text{ and } B > 0$$

As an exercise in the homework, express k in terms of α and B and show that the mean

$\mu = \alpha^{-1/B}\, \Gamma\left(1 + \frac{1}{B}\right)$.

Beta Distribution

The beta distribution is used as a random variable to approximate percentages of malfunctioning parts and other difficult estimations. Suppose the random variable is the time to repair some failed piece of equipment. The first step is to estimate the smallest possible repair time (x = a) and the largest possible repair time (x = b). We want P(x < b) = 1. We then want to create an appropriate probability density function on [a, b]. The beta distribution uses two parameters α_1

and α_2. It offers great flexibility because of the many different possibilities that the density function can assume by adjusting α_1 and α_2. The density becomes the uniform density if $\alpha_1 = \alpha_2 = 1$. This model is particularly useful if you have little knowledge about the random variable under consideration. It is recommended for most real world applications that the density function be skewed to the right. This density requires that $\alpha_2, \alpha_1 > 0$. Such modeling problems benefit from the many graphs that the beta function has depending on the values of α_1 and α_2. The density function is written:

$$f(x) = \frac{x^{\alpha - 1} (1 - x)^{B-1}}{B(\alpha, B)} \qquad \begin{matrix} 0 < x < 1 \\ \alpha, B > 0 \end{matrix}$$

$$B(\alpha,B) = \int_0^1 \mu^{\alpha - 1} (1 - \mu)^{B-1} \, du$$

$B(\alpha,B)$ is called the Beta Function. The mean of the beta function $\mu = \frac{\alpha}{\alpha + B}$; its variance $\sigma^2 = \frac{\alpha B}{(\alpha + B)^2 (\alpha + B + 1)}$

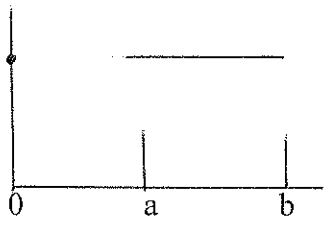
For a comprehensive treatment of the possible graphs of the beta distribution, please refer to the earlier cited text by Law and Kelton, p. 338-339.

Graphs

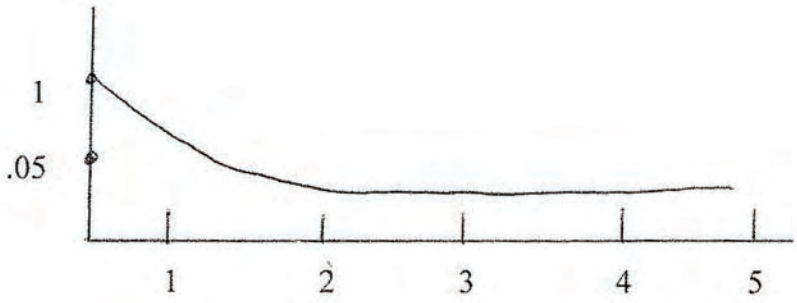
We conclude this section with graphs of certain continuous probability distributions.

Uniform

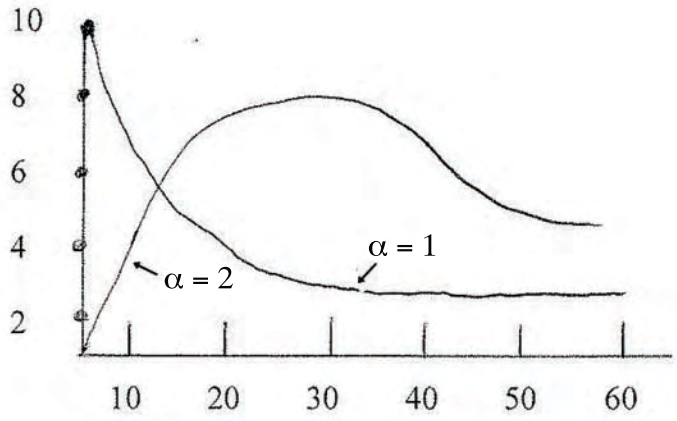
$$f(x) = \frac{1}{b - a}$$



Exponential

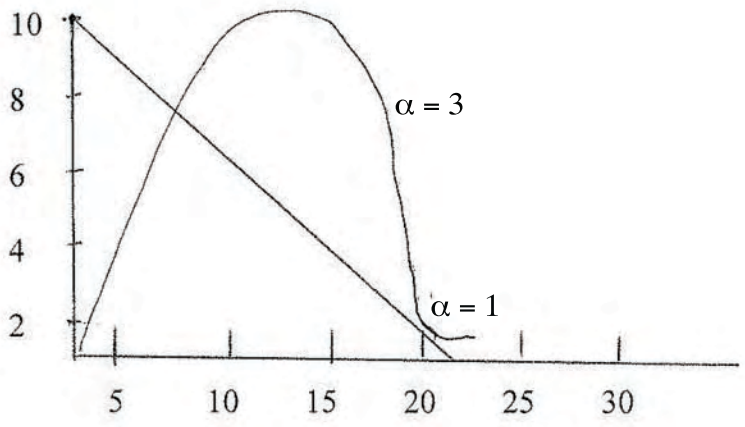


Gamma



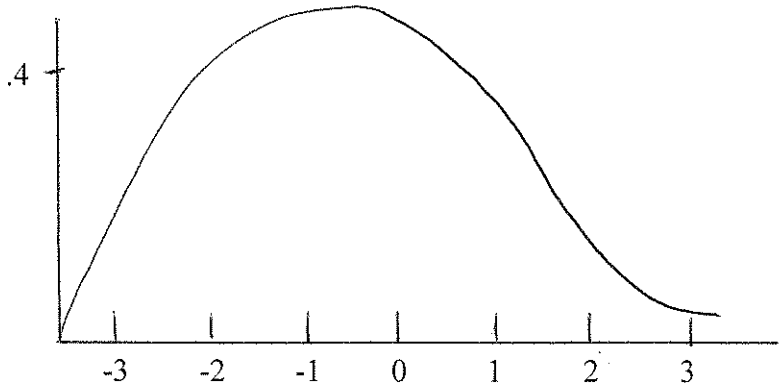
Weibull (α, 1)

Density



Normal Density

N (0, 1)



EXERCISES 3.2

1. Calculate the probability of two consecutive aces (without replacement) using the hypergeometric density.

2. Show that $\Gamma(k+1) = k!$ for $k = 0, 1, 2, \ldots$

3. If there are 10 red balls, 5 white balls, and 15 green balls in a box, calculate the probability that in 10 trials you choose 5 red, 3 white, and 2 green balls.

4. Graph the Weibull $(\alpha, 1)$ density if:

   a) $\alpha = 2$

   b) $\alpha = \frac{1}{2}$

5. a) Show that k can be expressed in terms of α and B for the Weibull distribution.

   b) Express the mean of the Weibull density in terms of the gamma function

   $\Gamma(1 + 1/B), \alpha, B > 0$

6. Show why numerical methods are necessary to obtain $F(x)$ from the normal density function.

3.3 Histograms

A graphical representation of data - the ***histogram*** - might give you an idea of which probability distribution to use in a problem. To create a histogram, follow the steps that are listed :

1) Start with the Range

$$\text{Range} = \text{Highest Value} - \text{Lowest Value}$$

2) Divide by k to obtain k intervals. Each interval has width = $\frac{\text{range}}{k}$

3) Count how many elements of data are in each interval.

4) Draw the histogram.

For example, let the arrival time in a bank be recorded. The times 8 AM = 1 and 4 PM = 8. Let n = 55. Suppose the 55 times are:

.0, .0, .0, .1, .1, .2, .3, .7, .9, .9, 1.0, 1.1, 1.1, 1.3

1.5, 1.6. 1.7, 2.0, 2.1, 2.2, 2.6, 2.8, 3.0, 3.1, 3.2, 3.2

3.3, 3.3 3.4, 3.6, 3.7, 4.0, 4.1, 4.2, 4.3, 5.1, 5.2, 5.3

5.4, 5.6, 6.0, 6.0, 6.1, 6.3, 6.4, 6.6, 6.7, 6.8, 7.0, 7.1

7.2, 7.3, 7.6, 7.8, 8.0

Now let us complete the histogram.

1) Range = 8.0 - .0 = 8

2) The number of categories are arbitrary. For computational ease, let k = 8. We also could have used Sturge's rule where k, the number of intervals can be selected:

$$k = 1 + \log_2 n = 1 + \log_2 55 = 6.8 \text{ or } 7.$$

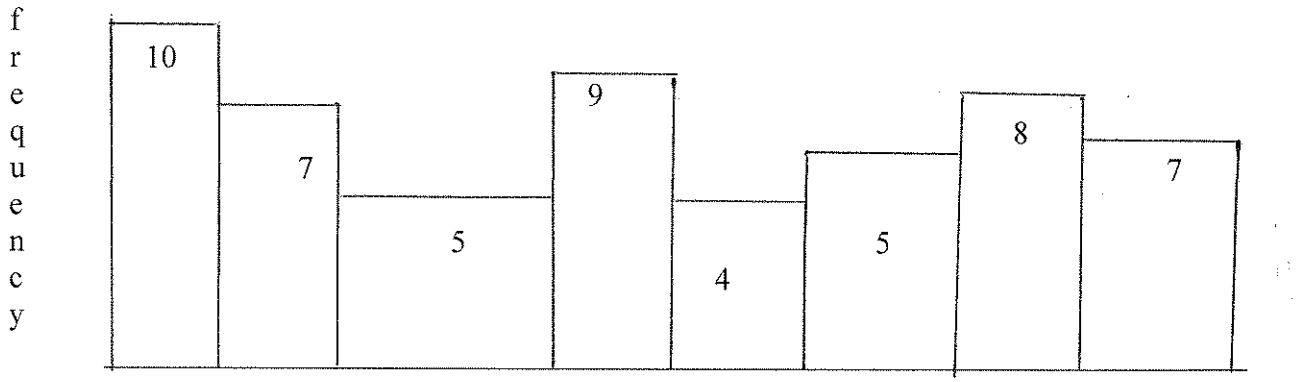
As Law and Kelton noted, several different values of k should be used in order to best approximate the density function of one of the major distribution functions.

interval = Δw = 8/8 = 1

3) Now create a table of frequencies.

interval	frequency
0 - .99	10
1 - 1.99	7
2 - 2.99	5
3 - 3.99	9
4 - 4.99	4
5 - 5.99	5
6 - 6.99	8
7 - 7.99	7

4) Now graph the results. The graph is called a ***histogram***.



You can look at the resulting histogram and take an educated guess as to the nature of the probability distribution. This one looks like a uniform density, since the times are fairly uniform.

However, we need the material from Section 3.5 - Goodness of Fit Analyses - to determine whether the distribution is what we believe it to be.

EXERCISES 3.3

1. For the previous example, let k = 4. Draw a histogram for this set of frequencies.

2. Airplane arrival times are listed in the following chart (noon - midnight). Draw a histogram to graphically represent the arrival times. 12 noon = 0, 12 midnight = 12. (n = 60)

.0, .0, .1, .3, .6, .7, .8, .9, 1.0, 1.1, 1.2

1.3, 1.3, 1.6, 1.9, 2.0, 2.1, 2.3, 2.4, 2.5, 2.6,

2.8, 2.9, 3.0, 3.2, 3.3, 3.4, 3.6, 3.7, 4.1, 4.1,

4.1, 4.3, 4.6, 4.7, 5.3, 5.8, 6.2, 6.8, 7.3, 7.3,

7.8, 8.0, 8.3, 8.4, 8.6, 9.3, 9.5, 9.7, 10.2, 10.2,

10.4, 10.6, 10.8, 10.9, 11.2, 11.4, 11.6, 12.0

3.4 Maximum Likelihood Estimates

As a final step toward selecting a probability distribution to fit real world data, we need a method to decide how sample data relates to the probability distribution in consideration. For example, if we hypothesize that a dice roll is a binomial distribution, we may with to estimate p - the probability of a seven. But perhaps the dice are loaded, and we don't know the true probability of seven. We simply could observe p - the proportion of sevens in n trials, and this could serve as our ***maximum likelihood estimate***. Sometimes ***maximum likelihood estimates*** of parameters agree with our intuition. For example, if we had a random sample of 100 U.S. citizens and asked them the number of hours each watched television weekly, the sample mean x would be used for a ***maximum likelihood estimate*** for the population mean of the national average for television viewing. Of course, advanced students pursuing their graduate degrees in mathematics (like you) will bring down this average. Since there are many instances where we can't easily determine an appropriate maximum likelihood estimate, we need a systematic procedure.

Each possible probability distribution that we could use to fit real world data has parameters. To illustrate, the normal density has two parameters, μ and σ. We will soon show how to take sample results and estimate parameters such as μ for the normal density. Since μ stands for the population mean, it makes sense to use $\bar{x}$, the sample mean, to represent the population mean. However, common sense can occasionally lead us to incorrect mathematical results. We need a procedure to arrive at these maximum likelihood estimates of the parameters associated with probability distributions. The method is based on elementary and intermediate calculus and is summarized as follows:

Steps to Obtain Maximum Likelihood Estimates of Parameters of Interest

1) Consider n observations $x_1, x_2, \ldots x_n$ from the probability distribution f(x).

2) Since the variables are assumed to be independent, we can create the ***likelihood function*** L. $L = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n)$.

3) Take the natural log of both sides of the equation.

$$\ln L = \ln f(x_1) + \ldots + \ln f(x_n)$$

4) Take the partial derivative of $\ln L$ with respect to the parameter of interest, say σ.

$$\frac{\delta \ln L}{\delta \sigma} = \frac{1}{f(x_1)} \frac{\delta f(x_1)}{\delta \sigma} + \ldots + \frac{1}{f(x_n)} \frac{\delta f(x_n)}{\delta \sigma}$$

5) Set $\frac{\delta \ln L}{\delta \sigma} = 0$ and solve for σ in terms of $x_1, x_2, \ldots x_n$.

This is the maximum likelihood estimate, $\hat{\sigma}$, for probability distribution parameter σ in terms of $x_1, x_2, \ldots x_n$.

For example, consider the Poisson distribution:

$$f(x) = \frac{\lambda^x e^{-1}}{x!} \quad x = 0, 1, 2, \ldots$$

Suppose $x_1, x_2, \ldots x_n$ are n observed values for x. Follow the steps to determine the maximum likelihood estimate for λ.

1) We are given values $x_1, x_2, \ldots x_n$

2) $L = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n)$

$$= \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdot ,,,, \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

$$= \frac{\lambda_{i=1}^{\sum^{n} x_i} \quad e^{-n\lambda}}{x_1!\ x_2!\ldots x_n!}$$

3) $\ln L = (\sum_{i=1}^{n} x_i)\ \ln \lambda + (-n\lambda) - \ln x_1! - \ldots - \ln x_n!$

4) $\frac{\delta \ln L}{\delta \lambda} = \sum_{i=1}^{n} x_i \left[\frac{1}{\lambda}\right] - n$

5) $\sum_{i=1}^{n} x_i \left[\frac{1}{\lambda}\right] - n = 0$

$$\sum_{i=1}^{n} x_i = n\lambda \rightarrow \lambda = \frac{\sum x_i}{n}$$

We ask the reader [Exercise 3.4, #1] to take the second partial with respect to L to show that $\frac{\delta^2 L}{\delta \lambda^2} < 0$. Therefore, $\frac{\sum x_i}{n}$ maximizes L.

Therefore, we can substitute the mean of observed x values as λ in a Poisson distribution. Then, as the next section will explain, we can test whether the Poisson distribution is a suitable model for the real world data.

Next, consider the data that is discrete and believed to be modeled by a geometric distribution. $f(x) = p\ (1-p)^x$, $x = 0, 1, 2, \ldots$, with $0 < p < 1$. We want to derive the maximum likelihood estimate for p from the sample data.

1) $x_1, x_2, \ldots x_n$ are observed values of x - the number of trials before the first success.

2) $L = p^n\ (1 - p)^{\sum_{i=1}^{n} x_i}$

3) $\ln L = n \ln p + \left[\sum_{i=1}^{n} x_i \right] \ln (1 - p)$

4) $\frac{\delta \ln L}{\delta p} = n \frac{1}{p} + \left(\sum_{i=1}^{n} x_i \right) \left[\frac{-1}{1-p} \right]$

$$\frac{\delta \ln L}{\delta p} = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i}{1-p}$$

5) $$\frac{n}{p} - \frac{\sum_{i=1}^{n} x_i}{1-p} = 0$$

$$\rightarrow p = \left[\frac{1}{\sum_{i=1}^{n} \frac{x_i}{n} + 1} \right] = \frac{1}{x + 1}$$

We leave it as a second exercise (3.4, #2) to show that $p = \frac{1}{\bar{x} + 1}$ maximizes L.

EXERCISES 3.4

1. Show that $\hat{\lambda} = \frac{\sum x_j}{n}$ maximizes L for the Poisson distribution.

2. Show that $\hat{p} = \frac{1}{x+1}$ maximizes L for the geometric distribution.

3. For the normal density function $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}$

show that the maximum likelihood estimate of:

a) the mean $\mu = \sum_{i=1}^{n} x_i / n$ [Assume σ^2 is known.]

b) the variance $\hat{\sigma}^2 = \frac{\sum_{j=1}^{n} (x_1 - \mu)^2}{n}$ [Assume μ is known.]

4) Show that the maximum likelihood estimater of p for a binomial density is $\hat{p} = \frac{\tilde{x}}{s}$, x is the number of successes in s independent trials. Show that $\hat{p} = \frac{\bar{x}}{s}$ maximizes L.

5) Show that $\hat{\mu} = \frac{\sum_{i=1}^{n} \ln x_i}{n}$ is a maximum likelihood estimator for μ in the lognormal density:

$$f(x) = \frac{1}{x\sqrt{2\pi}\,\sigma} e^{-(\ln x-\mu)^2/2\sigma^2} \quad x > 0$$

$$= 0 \quad \text{elsewhere}$$

3.5 Goodness of Fit Tests

The next step in creating models of real world phenomena is testing whether the data we observe fits hypothesized probability distributions. Rarely, if ever, do the data conform perfectly to any probability distribution. What we want is a good fit between the data and the probability distribution that will eventually model the data.

We will consider two goodness of fit tests in this section. The first - a χ^2 (***chi-square test***) - compares the observed scores with expected scores assuming a hypothesized probability distribution. The second - a ***Kolmogorow-Smirnov goodness of fit test*** - compares the hypothesized probability distribution to the empirical sample distribution.

To illustrate the chi-square test, consider a coin flip. Assume that you flip a coin 140 times and obtain 80 heads. Could you conclude that the coin was fair? Let $\alpha = .05$.

The level of Type I error, $\alpha = .05$, means that there is a 5% chance of rejecting the null hypothesis even if the null hypothesis is correct.

1) Use the formula $\chi^2 = \dfrac{\sum (\text{observed} - \text{expected})^2}{\text{expected}}$

2) Draw a table.

	Heads	Tails
Obs.	80	60
Exp.	70	70

The only non-trivial part of the chi-square is obtaining the expected. We assume H_o: $p_1 = p_2 = .50$ in hypothesizing the coin is fair. If the coin is fair, expected number of heads $= .5\ (140) = 70$.

3) Next compute $\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(80 - 70)^2}{50} + \frac{(60 - 70)^2}{50}$

Computed $\chi^2 = 4$.

4) Next, look up critical χ^2 in a table.

The larger the value of computed χ^2, the worse the fit is between the observed data and the hypothesized probability distribution. In this case, the hypothesized probability distribution was a binomial with P(head) = P(tail) = 1/2.

The test requires that the expected frequencies for each of the k cells be greater than or equal to 5. This could be expressed as $n p_i \geq 5$ for $i = 1, 2, \ldots, k$. The degrees of freedom, v, for the chi-square test are: (k = number of categories)

a) $v = k - 1$ if we don't use sample statistics to estimate population parameters.

b) $v = k - 1 - r$ if r population parameters have been estimated from sample statistics.

For example, if we use a sample mean to estimate a population mean and this is the only population parameter that has been estimated, then $r = 1$. For performing a goodness of fit test for a normal density, we have to estimate both μ and σ. Therefore, $r = 2$ for this case.

5) For our example, $k = 2$, $v = 1$ df

Critical x^2 .05, 1 = 3.84

6) Conclusion: $4 > 3.84$

At the $\alpha = .05$ level, we conclude $p_1 \neq p_2$. The coin is not fair.

A note of caution: If you go to the Taj Mahal casino, don't complain to the house if red comes out 80 out of 140 times over the course of the evening. Over 20 sequences of 140 trials, we

can expect to reject the null hypothesis once at the .05 level. And even if we monitored the table for 140 trials with the previous result, we can only be 95% confident of our result that $p_1 \neq p_2$.

The ***chi-square distribution*** can be derived as follows:

Let $x_1, x_2, \ldots, x_v$ be v independent normally distributed random variables having mean $\mu = 0$ and variance $\sigma^2 = 1$. Consider the following random variable χ^2:

$$\chi^2 = x_1^2 + x_2^2 + \ldots + x_v^2$$

For $x \geq 0$,

$$P(\chi^2 \leq x) = \frac{1}{2^{v/2}\,\Gamma(v/2)} \int_0^x u^{(v/2)-1}\, e^{-u/2}\, du \quad \text{if } x > 0$$

$$P(\chi^2 \leq x) = 0 \quad \text{if } x \leq 0.$$

This distribution function leads to the critical chi-square value that we find in statistical tables. The chi-square distribution is a special case of the gamma distribution if $\alpha = v/2$ and $B = 2$.

For a second example, let us use the chi-square goodness of fit test to determine whether the digits of an irrational number are uniformly distributed. We will discuss further the notion of uniformity in the chapter on random numbers.

Consider the table below as the frequency of 0, 1, 2, …, 9 that occur in a 1000 digit approximation to the irrational number $\sqrt{x}$ (where x is not a perfect square).

	0	1	2	3	4	5	6	7	8	9
Observed	95	103	102	90	115	94	103	98	106	94
Expected	100	100	100	100	100	100	100	100	100	100

The observed are simply the frequency of digits that a computer approximation of $\sqrt{x}$ has counted. The expected are obtained by use of the assumption that the digits are uniform. Therefore, $P(1) = P(2) = \ldots = P(9) = 1/10$.

The expected for each category are calculated by multiplying 1/10 by total v. This yields $1/10 \cdot (1000) = 100$ for each category.

Use $\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(95 - 100)^2}{100} + \frac{(103 - 100)^2}{100} + \ldots + \frac{(94 - 100)^2}{100} = 4.84$

Degrees of freedom = 10 - 1 = 9

CRITICAL χ^2 .05, 16.9

Therefore, accept H_o: $P(1) = P(2) = \ldots = P(9) = 1/10$

The digits of $\sqrt{x}$ are uniformly distributed.

Kolmogorov-Smirnov Test

The chi-square goodness of fit test has several disadvantages. The number of intervals is arbitrary; for example, for $\sqrt{x}$ we could have selected 100 intervals (00, 01, . . . 99) with $P(00) = \ldots = P(99) = 1/100$. Also, the test is valid only asymptotically - that is, as $n \rightarrow \infty$.

For the Kolmogorov-Smirnov (K-S) test, the intervals are not arbitrary, there is no need to group data, and the method is valid for any n. The key to the validity of the method is that if x is a continuous random variable with distribution function F(x), F(x) is uniformly distributed on (0,1). The proof is presented in great detail in *Introduction to Probability Theory* by Hoel, Port and Stone (Houghton Mifflin, 1971), pp. 119-120.

The Kolmogorov-Smirnov test compares the empirical (sample) distribution with the hypothesized probability distribution. The empirical distribution $s_n(x)$ is defined as follows:

1) Arrange the sample in numerical order

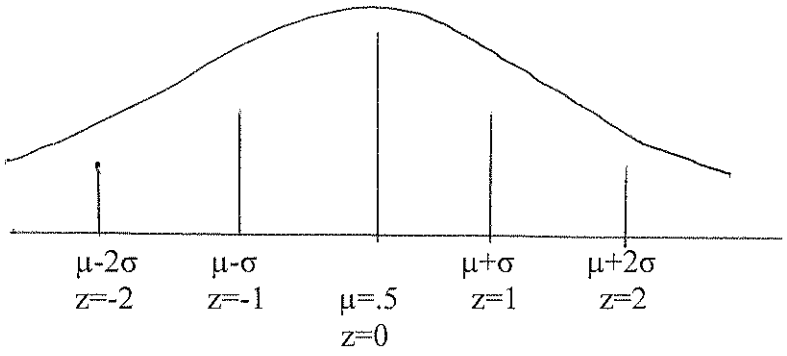
$x_1 \leq x_2 \leq \ldots \leq x_n$

2) Define $s_n(x) = 0$ if $x < x_1$

j/n $x_j \leq x < x_{j+1}$

1 $x_n \leq x$

For the following sample of twenty data values from a distribution that we hypothesized to be a normal distribution, N(0,1), $\mu = .5$, $\sigma = .16$.

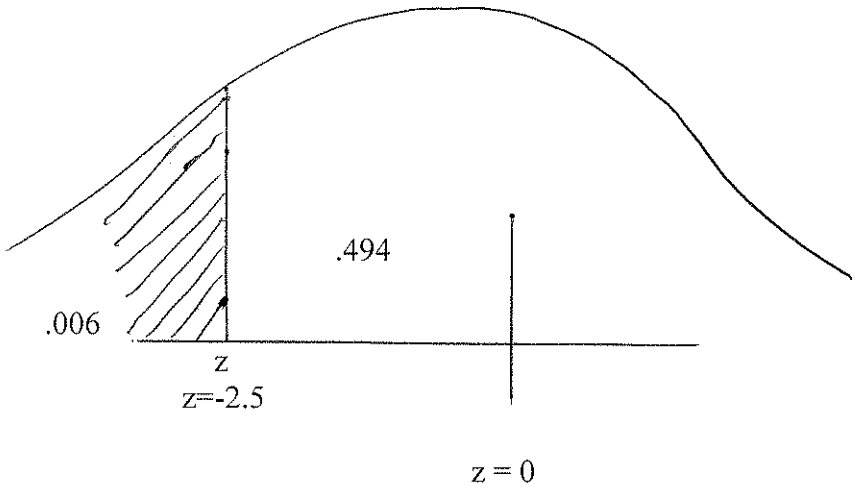


Let $x_1, x_2, \ldots, x_{20} = 0, 0, .1, .1, .2, .3, .4, .4, .5, .5, .6, .6, .7, .7, .8, .9, 1.0, 1.0, 1.0, 1.0$.

j =	2	4	5	6	8	10	12	14	15	16	20
x_i	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
$F(x_i)$	.01	.01	.03	.11	.26	.50	.74	.90	.97	1.0	1.0
$s_n(x_i)$	.1	.2	.25	.3	.4	.5	.6	.7	.75	.8	1.0

Note we use the maximum value of j corresponding to each element x_i in defining $s_n(x) = j/n$. For example there are two values for $x_i = 0$, both x_1 and x_2. The variable $s_n(x_i)$ is calculating the cumulative relative frequency of x_i - that is, the percentage of values less than or equal to x_i. For example, $s_n(0) = 2/20 = .1$ since 0 has frequence of 2.

To compute $F(x_i)$ we need to find a z score for each x_i and use a table of standard normal values to look up the probability values corresponding to each z score. To compute the z score for 0, $z = \frac{0 - .5}{.16} = -3.125$. The P value corresponding to $z = -3.125$ is .001. This is the value of F(0). The graph which you should always consult to understand the meaning of the standard normal values, is below. The shaded value is F(z).



To calculate F(.1),

1) $z = \frac{.1 - .5}{.16} = -2.5$

2) $P(z \leq -2.5) = .5 - .494$

$P = .006$

The calculation of F(.2), F(.3), . . . , F(1) is left to the reader.

To proceed with the K-S test, we use two values - the maximum value of $|F(x_i) - s_n(x_i)|$ and the maximum value of $|F(x_i) - s_n(x_{i-1})|$. These values are computed in a table of D_n values, where $D_n = \sup |F(x) - s_n(x)|$. The D_α table gives you critical values for the Kolmogorov-Smirnov test. In our example, $\max |F(x_i) - s_n(x_i)| = .22$.

$F(.2) - s_n(.2) = .22$; also $F(.8) - s_n(.8) = .22$

Next, max $|F(x_i) - s_n(x_{i-1})|$ can be calculated by looking at the diagonals of the table, $|F(.1) - s_n(0)|$, $|F(.2) - s_n(.1)|$, etc.

$$\max |F(x_i) - s_n(x_{i-1})| = |F(.7) - s_n(.6)| = .30$$

Take the maximum of .22 and .30, which is .30 and look up the critical value of D_α in a table. For $n = 20$, $\alpha = .05$, $D_\alpha = .29$. Our computed value $D_{max} = .30$. Therefore, at $\alpha = .05$, we conclude that our sample is not from a normal density with mean .5, $\sigma = .16$. However, if we adopted the marginal $\alpha = .10$ level, critical $D_{.10} = .26$. So, at the $\alpha = .10$, we could conclude that our sample was from our hypothesized normal density. Generally, we look for $\alpha = .05$ as our standard level of Type I error.

The K-S test has several limitations. According to Law and Kelton (1991), these include the facts that:

1) For discrete data, the critical values are not easily obtainable. Numerical procedures can be performed fairly easily only if $n \leq 50$.

2) The original form of the K-S test is valid only if all of the parameters of the hypothesized distribution are known and the distribution is continuous.

3) The K-S test - in its original all-parameters-known form - has frequently been applied to any continuous distribution with estimated parameters and for discrete distributions. However, the probability of a Type I error will be smaller than that specified.*

---

*Law, A. and Kelton, W. *Simulation Modeling and Analysis* (McGraw Hill, New York, 1991), pp. 387-389.

EXERCISES 3.5

1. During an experiment for ESP, the roller of a die was asked to roll sixes. He rolled 250 sixes out of 1620 trials. Was his performance above the chance level ($\alpha = .05$)?

2. For the following sample of 50 observations, test whether the distribution is:

   a) uniform

   b) normal [use both chi-square and K-S tests with $\alpha = .05$]

0, 0, .1, .1, .1, .2, .2, .2, .3, .3, .4, .4, .4, .5, .5, .5, .5, .5, .5, .5, .6, .6, .7, .7, .8, .9, 1.0, 1.0, 1.1, 1.2, 1.2, 1.3, 1.3, 1.3, 1.4, 1.4, 1.4, 1.4, 1.5, 1.5, 1.6, 1.6, 1.7, 1.8, 1.9, 2.0, 2.0, 2.0

3) How many times out of 1000 would the dealer have to win in order to conclude that the results were greater than chance at a game where the house advantage is 52% - 48%? ($\alpha = .05$)

4) In a bank the following times in minutes were observed for tellers to complete transactions. Determine which probability distribution is a good fit. Use maximum likelihood estimation together with the techniques from this section.

1.01	2.33	1.15	1.23	2.35	5.12	.35
2.15	1.36	1.36	1.11	1.36	2.15	2.18
2.11	2.19	1.73	3.19	1.15	5.11	2.56
2.19	3.11	.13	1.37	3.15	2.46	1.63
1.38	1.21	1.23	2.19	2.39	1.23	1.41
1.96	1.56	2.15	2.36	2.32		

5) Use SPSS

For the following service times, test whether the probability distribution is:

a) Exponential

b) Uniform

c) Normal

For the uniform test, use SPSS and separately use the chi-square goodness of fit tests after using SPSS to create 4 categories, 25%ile, 50%ile, 75%ile, and 99+%ile.

Instructions

SPSS uses the K-S goodness of fit test

Steps

1) Go to File

2) Enter 48 numbers in VAROOO1 column

3) Analyze

4) Non-parametric

5) Legacy Dialogs

6) 1 Sample K-S test

7) Click VAR 0001 into test variable list

8) Click normal, uniform, and exponential

9) OK

10) The K-S test uses H_o: the density is a good fit. H_a: the density is not a good fit

If the asymptotic significance level < .05, reject H_o. The data is not a good fit to the hypothesized density. Goodness of fit is a poor name. The test only rejects very poor fits.

For the chi-square test of uniformity, follow these steps:

1) VAR 0001 - entered data

2) Analyze

3) Descriptive statistics

4) Frequencies

5) Check VAR 0001 to variable(s)

6) Statistics

7) Cut points to 4 = groups

8) Use the 25%, 50%, 75% and 99%ile to identify the number of observed. Use the interval to calculate the expected. Since the data range over (0,10), use (.251 - 0) = .251 as probability to compute the expected as follows:

$(.251)\ (48) = 12 = l_1$

$(.45 - .25)\ (48) = 9.6 = l_2$

Find l_3 and l_4 and use $X^2 = \Sigma \frac{(0 - E)^2}{E}$

2.20	8.50	6.10	1.18	2.66	6.16	2.90
2.60	.88	2.80	5.03	2.72	7.47	4.50
9.45	7.89	3.18	6.13	1.30	9.87	6.60
7.10	2.93	1.77	5.72	2.51	3.06	6.46
2.74	7.43	4.45	9.31	.06	6.27	8.89
4.80	.81	8.46	6.12	1.20	9.90	8.33
8.93	1.96	.04	3.48	7.58	1.63	

SPSS ACTIVITY FOR GOODNESS OF FIT TESTS

Goodness of fit tests with SPSS analyze data for the important and widely used continuous normal, exponential, and uniform densities as well as the discrete Poisson density.

Let us open the 49 variable data set and use SPSS to analyze goodness of fit of the data to the above densities.

We have already performed correlational analyses and revealed significant findings. Each statistical test is based on proofs that are derived from the calculus. Each proof requires assumptions. Correlation assumes that the pairs (x, y) has a bivariate normal distribution. The testing of such an assumption is an advanced exercise in Numerical Analysis, since the normal density function $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2(x-\mu/\sigma)^2}$ has no simple integral to measure probability. The bivariate normal density is even more complex. Please refer to Burden and Faires, Numerical Analysis, PWS Kent for a thorough explanation of how to find the probabilities of these densities, enabling the researcher to perform a goodness of fit test.

On a practical level, sample sizes are usually too small to allow a bivariate density goodness of fit test. We could settle for analyzing whether for any fixed value of x, the corresponding values of y are normally distributed.

Let us use SPSS to test whether selected variables are normally distributed.

Use the instructions from Problem 5 - service time goodness of fit tests to accept or reject the normal distribution as a good fit.

Note the output has .000 as the Asymptotic Significance Level for K-S goodness of fit test. This low level (< .05) means that the normal distribution is not a good fit. The K-S test uses the

following: H_0: the null hypothesis states that the distribution is a good fit. H_a: the two tailed alternative hypothesis is that the distribution is not a good fit.

One of the problems with goodness of fit tests is the tendency to accept distributions as good fits whenever they are not significantly departing from a hypothesized distribution. Also the chi-square test yields slightly different results, even when testing the same data as the K-S test. Advanced students could spend profitable hours researching the circumstances where the two goodness of fit tests converge or diverge, based on the data and properties of the exponential, normal, and uniform distributions.

Let us follow the same steps, click the boxes for uniform and exponential distributions and inspect the output. All three variables fail to qualify as uniform or exponential distributions as well.

As a possible follow-up activity the student is encouraged to collect 100 elements of data from the real world. Possible rich areas for exploration are grade distributions for a course such as Calculus (which may be normally distributed) or waiting times in queues in banks or stores (which may be exponentially distributed).

An interesting application of the Poisson distribution, that would help students learn the importance of this function, is to gather accident statistics from intersections that the Department of Transportation cites as dangerous. For example, this author consulted for a school district to analyze the safety of transporting students to a high school 28 miles away. Extensive research of the route led to finding two dangerous intersections with P(an accident) = 3/100,000 at each. If we simplified our model and combined the probabilities, we could estimate P(accident) $\approx$ 6/100,000 = 3/50,000.

During four years of high school, the bus would travel over these intersections $180 \cdot 8 = 1440$ times. The Poisson distribution has two parameters, x and λ. x refers to in this case the number of accidents; λ refers to n Θ from the binomial density and in this case $n = 1440$, $\Theta = 3/50{,}000$. $\lambda = 1440\ (3/50{,}000) = .0864$ P(0 accidents) $= \frac{\lambda^x e^{-\lambda}}{x!} = \frac{.0864^0\ e^{-.0864}}{0!} = .92$

$\therefore$ P(at least one accident) = 8%

We rejected the plan to bus our children 28 miles. The risk was too great.

Please obtain data from the real world and perform goodness of fit tests with SPSS (K-S test) and chi-square tests with the same data. Compare results. Explore situations where one test is preferable to the other, based upon the properties of the data set. This is a fertile research direction in Statistical Modeling.