NYU

ECON UA-18

Chapter 4 – Special Distributions

Dr. Ralph Chikhany

4.1, 4.2 – Introduction, Poisson and Exponential Distributions

Introduction: from Binomial to Poisson

Theorem 4.2.1 *Suppose X is a binomial random variable, where*

$$P(X = k) = p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

If $n \to \infty$ and $p \to 0$ in such a way that $\lambda = np$ remains constant, then

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np = \text{const.}}} P(X = k) = \lim_{\substack{n\to\infty \\ p\to 0 \\ np = \text{const.}}} \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-np}(np)^k}{k!}$$

If a typist averages one misspelling in every 3250 words, what are the chances that a 6000-word report is free of all such errors? Answer the question two ways— first, by using an exact binomial analysis, and second, by using a Poisson approximation.

The Poisson Distribution

Poisson

Given: a series of events occurring independently at a constant rate of λ per unit time (or volume or space)

$$\begin{aligned}
\text{Let } X &= \text{number of occurrences in unit time (or volume or space)} \\
p_X(k) &= P(k \text{ events occur in unit time (or volume or space))} \\
&= e^{-\lambda}\lambda^k/k!, \quad k = 0, 1, 2, \ldots \\
E(X) &= \lambda \\
\mathrm{Var}(X) &= \lambda
\end{aligned}$$

Suppose that on-the-job injuries in a textile mill occur at the rate of 0.1 per day. What is the expected number of injuries over the next (five-day) workweek? What is the probability that two accidents will occur during this same period?

From Poisson to Exponential

Situations sometimes arise where the time interval between consecutively occurring events is an important random variable. Imagine being responsible for the maintenance on a network of computers. Clearly, the number of technicians you would need to employ in order to be capable of responding to service calls in a timely fashion would be a function of the "waiting time" from one breakdown to another. Suppose a series of events satisfying the Poisson model are occurring at the rate of λ per unit time. Let the random variable Y denote the interval between consecutive events. Then Y has the exponential distribution.

Exponential

$$f_Y(y) = \lambda e^{-\lambda y}, y > 0$$

$$E(Y) = 1/\lambda$$

$$\mathrm{Var}(Y) = 1/\lambda^2$$

Example

The length of time for one individual to be served at a cafeteria is a random variable having an exponential distribution with a mean of 4 minutes. What is the probability that a person is served in less than 3 minutes?

Records show that deaths occur at the rate of 0.14 per day among patients residing in a large nursing home. If someone dies today, what are the chances that a week or more will elapse before another death occurs?

4.3 – 4.6 – [Some] More Distributions

The Negative Binomial Distribution

Given: n independent trials, each having probability p of success

$$\text{Let } X = \text{trial where } r\text{th success occurs}$$

$$p_X(k) = P(r\text{th success occurs on } k\text{th trial})$$

$$= \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, r+2, \ldots$$

$$E(X) = r/p$$

$$\text{Var}(X) = [r(1-p)]/p^2$$

A door-to-door salesperson is required to document five in-home visits each day. Suppose that she has a 30% chance of being invited into any given home, each one representing an independent trial. What is the probability she requires fewer than eight houses to achieve her fifth success?

The Geometric Distribution

Given: n independent trials, each having probability p of success

$$\begin{aligned} \text{Let } X &= \text{trial where first success occurs} \\ p_X(k) &= P(\text{First success occurs on } k\text{th trial}) \\ &= (1-p)^{k-1}p, \quad k = 1, 2, \ldots \\ E(X) &= 1/p \\ \text{Var}(X) &= (1-p)/p^2 \end{aligned}$$

You decide to keep placing a $1 bet on number 15 in consecutive spins of a roulette wheel until you win. On any spin, there's a 1-in-38 chance that the ball will land in the 15 slot. How many spins do you expect it to take until you win? What is the probability it takes you 5 spins to win?

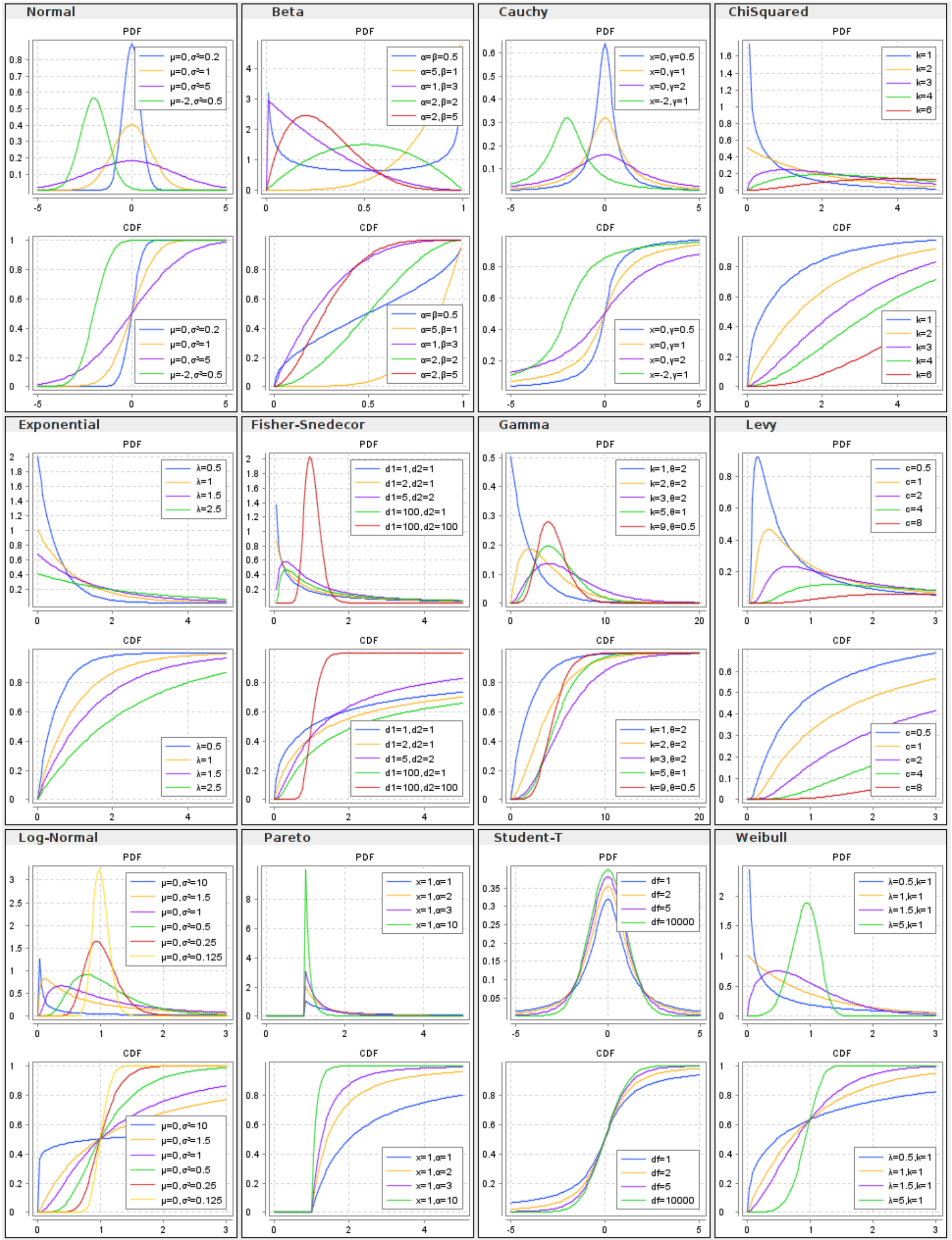
The Uniform Distribution

$$f_Y(y) = 1/\theta, 0 \leq y \leq \theta$$
$$E(Y) = \theta/2$$
$$\text{Var}(Y) = \theta^2/12$$

Suppose the time it takes a student to finish a quiz is uniformly distributed between 6 and 15 minutes, inclusive. Let *t* denote the time, in minutes, it takes a student to finish a quiz. Find the probability that a randomly selected student needs at least eight minutes to complete the quiz.

Other Distributions

Gamma Distribution
Beta Distribution
Weibull Distribution
Lognormal Distribution
and so many more!



Which One?

Before Recitation, prepare the following exercises. At the very least, try to decide which distribution to use if it is not given.

1) An airline sells 200 tickets for a certain flight on an airplane that has only 198 seats because, on the average, 1 percent of purchasers of airline tickets do not appear for the departure of their flight. Determine the probability that everyone who appears for the departure of this flight will have a seat.
2) When a player plays chess against their favorite computer program, they win with probability 0.60, loses with probability 0.10, and the rest of the times the games result is a draw.
   a. Find the probability that the player wins for the first time during their third game.
   b. Find the probability that the player's fifth win happens when they play their eighth game
3) An inventory control employee selects 2 RAM modules from a shipment of 6 known. Let X be the random variable for the number of defective modules from her selection. Create the probability distribution of X, and find the expected number of defective modules.
4) The interval, X seconds, between cars passing a point on a motorway follows an exponential distribution with mean ½.
   a. Find the pdf and cdf of X.
   b. Calculate the probability that the interval until the next car passes is longer than 3 seconds.
   c. State a distribution that could be used to model the number of cars passing the point each second, giving the values of any parameters.
5) Alex takes a multiple choice quiz in his Anthropology 100 class. The quiz has 10 questions, each has 4 possible answers, only one of which is correct. Alex did not study for the quiz, so he guesses independently on every question. What is the probability that Alex answers exactly 2 questions correctly?
6) An oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil.
   a. What is the probability that the first strike comes on the seventh well drilled?
   b. What is the mean and variance of the number of wells that must be drilled if the oil company wants to set up three producing wells?

4.3 – The Normal Distribution

Figure arrangement and some examples in this section are adapted from
Engineering Statistics byMontgomery, Runger and Hubele, 5th Ed, Copyright © 2011 Wiley

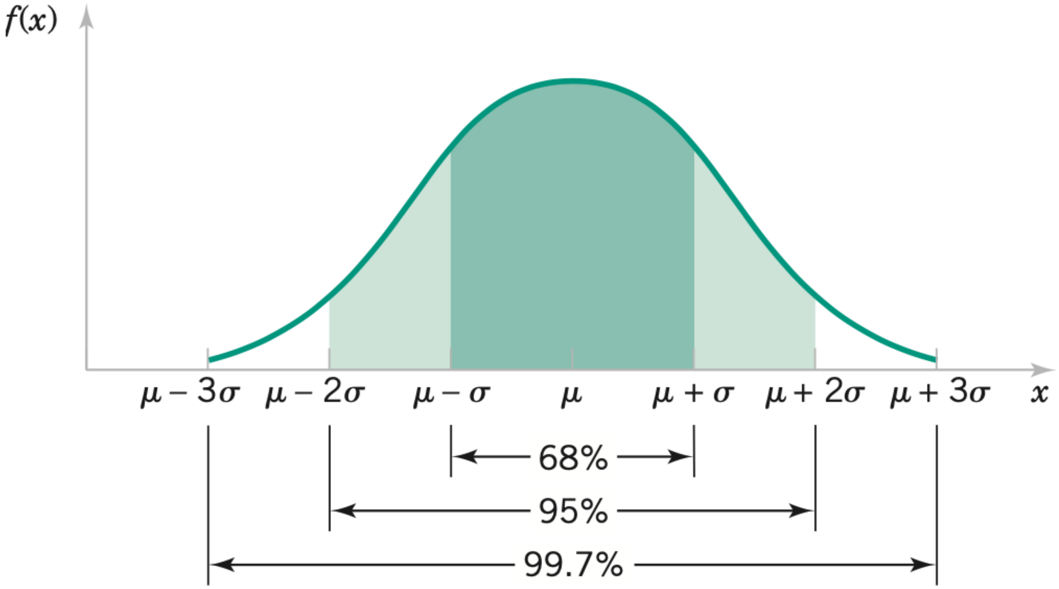
Introduction

Undoubtedly, the most widely used model for the distribution of a random variable is a **normal distribution**. You might have seen histograms with symmetric, bell shapes, these are reminiscent of this [continuous] distribution. A random variable with the probability density function on the right follows a normal distribution with parameters μ and σ where $\mu \in \mathbb{R}$ and $\sigma > 0$ are respectively the mean (expected value) and standard deviation σ.

Definition 4.3.1
A random variable Y is said to be normally distributed with mean μ and variance σ^2 if

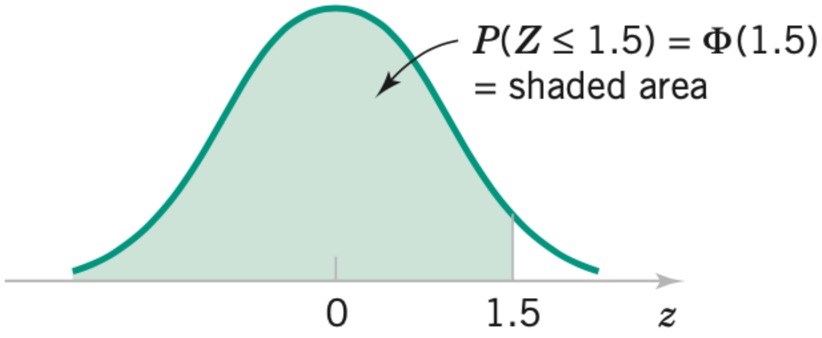
$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad -\infty < y < \infty$$

The symbol $Y \sim N(\mu, \sigma^2)$ will sometimes be used to denote the fact that Y has a normal distribution with mean μ and variance σ^2.



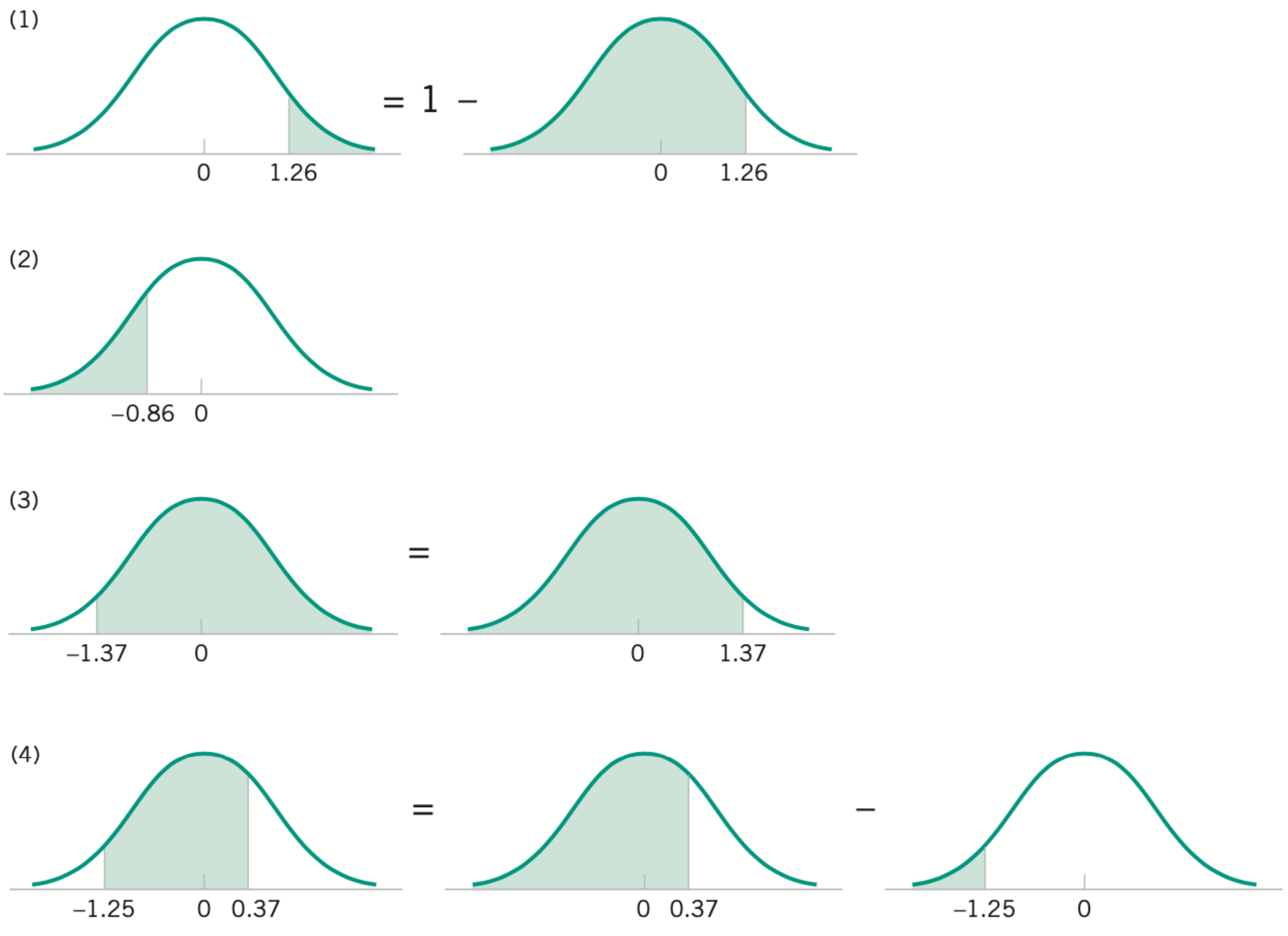
The Standard Normal

A normal random variable with $\mu = 0$ and $\sigma = 1$ is called a standard normal random variable, denoted by Z. Then $\Phi(z) = P(Z \le z)$ is used to denote the **cumulative distribution function** of a standard normal random variable. A table (or computer software) is required because the probability cannot be calculated by elementary methods of integration. Probabilities that are not of the form $P(Z \le z)$ are found by using the basic rules of probability and the symmetry of the normal distribution along with the normal table. The following examples illustrate the method.



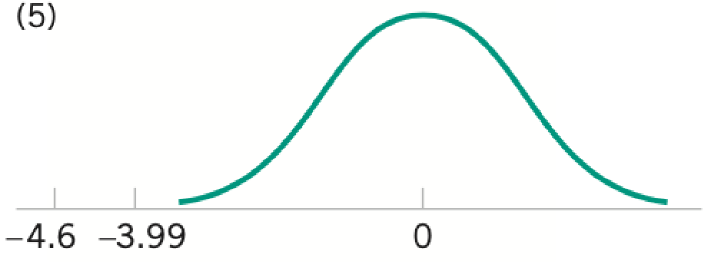
z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

More Examples

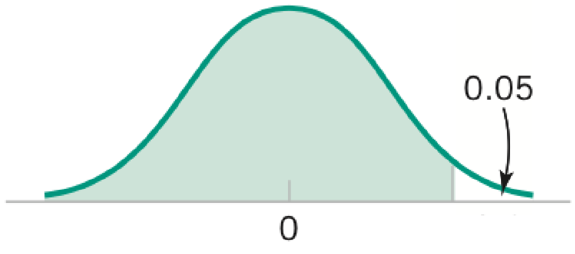


More Examples

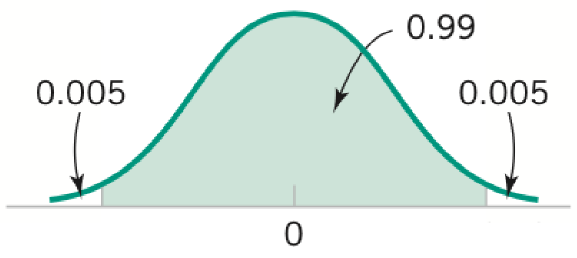
(5)



(6)



(7)



Standardizing Normal Random Variables

Typically, rvs that are normally distributed are not standard (meaning the mean and standard deviation are not 0 and 1). Thus, we can **standardize** a normally distributed random variable by using the following transformation:

$$P(a \le Y \le b) = P\left(\frac{a-\mu}{\sigma} \le \frac{Y-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

The ratio $\frac{Y-\mu}{\sigma}$ is often referred to as either a Z *transformation* or a Z *score*.

Example

The sales revenue generated by books signed to a specific publisher can be modeled by a normal distribution with a mean of \$6000 and a standard deviation of \$100.

(a) What is the probability that an author generates less than \$6250 in revenue?

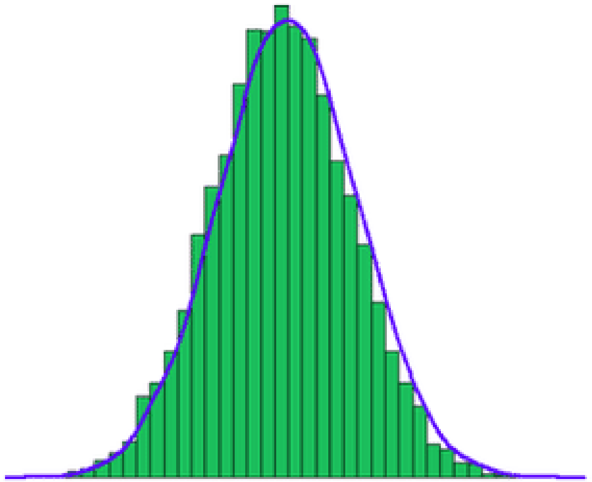
(b) What is the sales revenue threshold for authors signed to the publisher that would represent the top 5% of revenue generated?

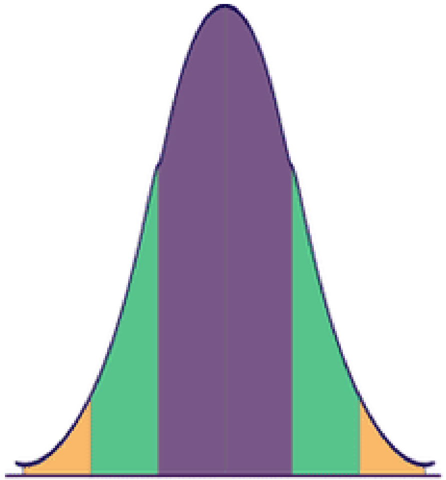
Examples

The fill volume of an automated filling machine used for filling cans of carbonated beverage is normally distributed with a mean of μ fluid ounces and a standard deviation of 0.1 fluid ounce. Suppose that μ is a parameter that can be adjusted easily. At what value should the mean be set so that 99.9% of all cans exceed 12 ounces?

The Normal Approximation to the Binomial

Suppose X is a [discrete] binomial random variable with n trials and probability of success p. Under certain conditions, X is approximately (but practically considered) normal with mean $\mu = np$ and variance $\sigma^2 = np(1 - p)$. In other words,

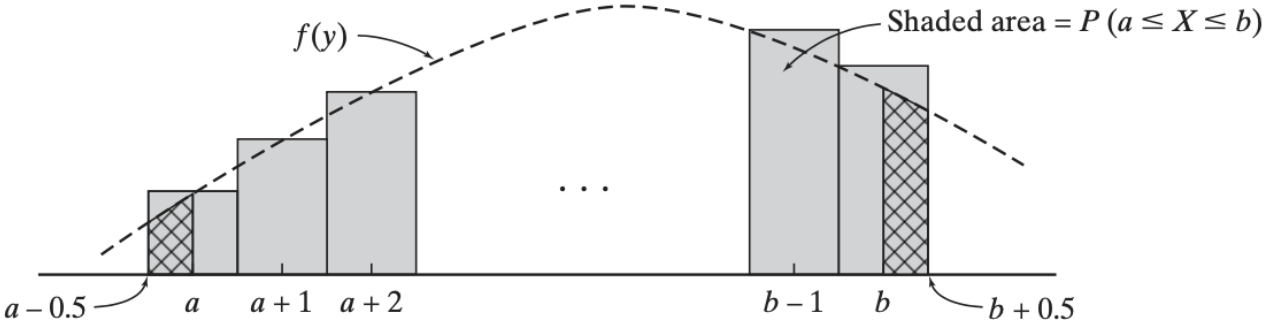




(DeMoivre-Laplace) Let X be a binomial random variable defined on n independent trials for which $p = P(\text{success})$. For any numbers a and b,

$$\lim_{n\to\infty} P\left[a \le \frac{X - np}{\sqrt{np(1-p)}} \le b\right] = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-z^2/2}\,dz$$

Conditions and Continuity Correction



Example

We've all had those annoying telemarketing calls from agents trying to sell us things we do not need. A call protection agency recently issued a report that suggested 45% of all cell phone calls this year will be spam. Suppose 500 cell phone calls are selected at random.

(a) Find the probability that at least 240 of the cell phone calls will be spam exactly and via an appropriate approximation.

(b) Find the probability that exactly 245 of the cell phone calls will be spam exactly and via an appropriate approximation.

Example

We’ve all had those annoying telemarketing calls from agents trying to sell us things we do not need. A call protection agency recently issued a report that suggested 45% of all cell phone calls this year will be spam. Suppose 500 cell phone calls are selected at random.

(a) Find the probability that at least 240 of the cell phone calls will be spam exactly and via an appropriate approximation.

(b) Find the probability that exactly 245 of the cell phone calls will be spam exactly and via an appropriate approximation.

Functions of Normal Random Variables

Recall: In general, $E(aX + bY) = aE(X) + bE(Y)$
but $V(aX + bY) \neq aV(X) + bV(Y)$ (see Chapter 3 slides)

Definition: A linear combination of n random variables $X_1, X_2 \cdots X_n$ is a new random variable defined by

$$Y = \sum_{i=1}^{n} c_i X_i$$

Results: In this case,

$$E(Y) = c_1 E(X_1) + c_2 E(X_2) + \cdots + c_n E(X_n).$$

If all the $X_i s$ are pairwise independent:

$$V(Y) = c_1^2 V(X_1) + c_2^2 V(X_2) + \cdots + c_n^2 V(X_n)$$

Functions of Normal Random Variables

$$\text{If } \{X_i\} \text{ are normal and independent } \Rightarrow Y = \sum_{i=1}^{n} c_i X_i \text{ is also normal}$$

Random Samples and Statistics

Results

The Central Limit Theorem

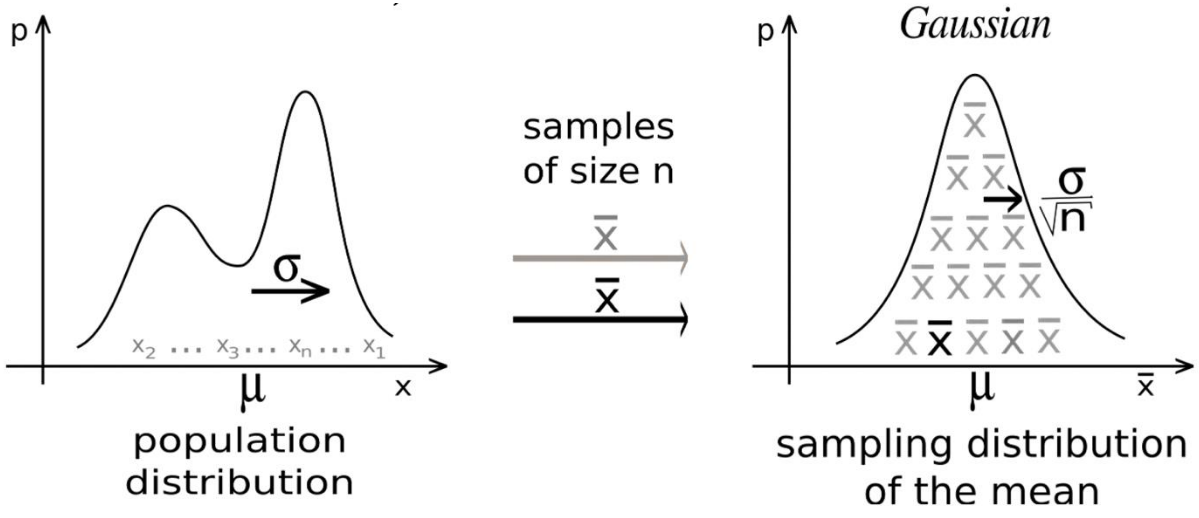
For large ($n \geq 30$) sample sizes, the random variable $\bar{x}$ has an approximate normal distribution, with mean and standard deviation

$$\mu_{\bar{X}} = \mu_X \qquad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

In other words, the random variable

$$Z = \frac{\bar{X} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$$

has an approximate standard normal distribution. This theorem holds regardless of the type of population distribution. The population distribution could be normal; it could be uniform (equally likely outcomes); it could be strongly positively skewed; it could be strongly negatively skewed. Regardless of the shape of the population distribution, the sampling distribution of the mean will be approximately normal for large sample sizes.



Examples

The time to complete a manual task in a manufacturing operation is a random variable with mean of 0.50 minute and standard deviation 0.05 minute. Find the probability that the average time to complete the manual task, after 49 repetitions, is less than 0.465 minute.

Papers produced by a certain machine are on average 10 micrometers thick, with a standard deviation of 1 micrometer. Assume that the thickness is normally distributed and that the thicknesses of different wafers are independent. Determine the number of papers that need to be measured such that the probability that the average thickness exceeds 11 micrometers is 0.01.Why is this sample size acceptable here?

CLT for Sums

Suppose that one hundred fair dice are tossed. Estimate the probability that the sum of the faces showing exceeds 370.