3. Continuous Random Variables

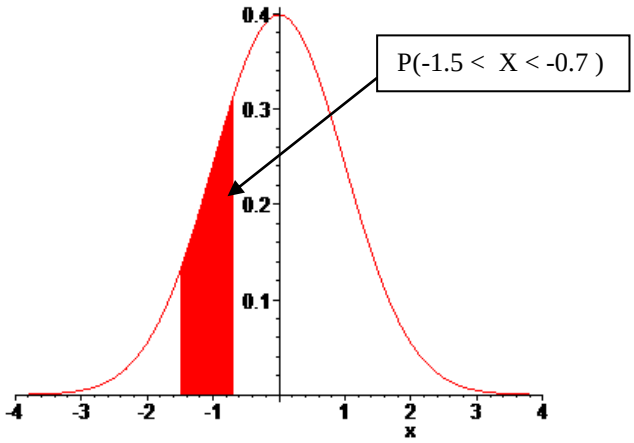
A continuous random variable is a random variable which can take values measured on a continuous scale e.g. weights, strengths, times or lengths.

For any pre-determined value *x*, P($X = x$) = 0, since if we measured *X* accurately enough, we are never going to hit the value x exactly. However the probability of some region of values near *x* can be non-zero.

Probability density function (pdf): $f(x)$

$$P(a \leq X \leq b) = \int_a^b f(x')dx'$$

Probability of X in the range a to b.



Since *X* has to have some value

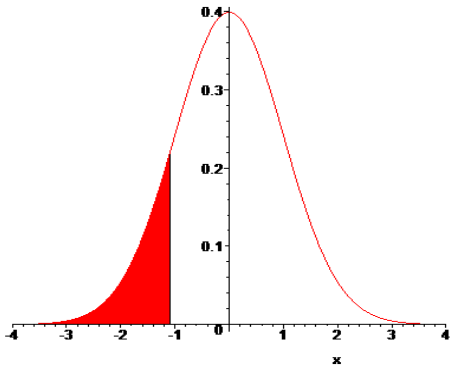
$$\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X < \infty) = 1$$

And since $0 \leq P \leq 1$, For a pdf, $f(x) \geq 0$ for all x.

Cumulative distribution function (cdf) :

This is the probability of $X < x$.

$$F(x) \equiv P(X < x) = \int_{-\infty}^{x} f(x')dx'$$



Mean and variance

Expected value (mean) of X: $\mu = \int_{-\infty}^{\infty} xf(x)dx$

Variance of X: $\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

Note that the mean and variance may not be well defined for distributions with broad tails. The *mode* is the value of x where $f(x)$ is maximum (which may not be unique). The *median* is given by the value of *x* where

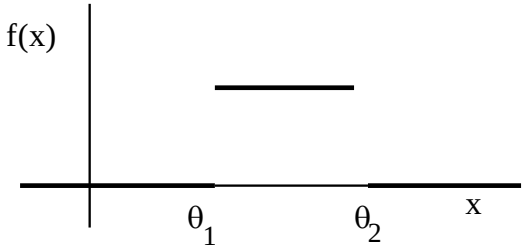
$$\int_{-\infty}^{x} f(x')dx' = \frac{1}{2}.$$

.

Uniform distribution

The continuous random variable X has the Uniform distribution between θ_1 and θ_2, with $\theta_1 < \theta_2$ if

$$f(x) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \theta_1 \le x \le \theta_2 \\ 0 & \text{otherwise} \end{cases}$$



$X \sim U(\theta_1, \theta_2)$, for short.

Roughly speaking, $X \sim U(\theta_1, \theta_2)$, if X can only take values between θ_1 and θ_2, and any value of X within these values is as likely as any other value.

Mean and variance: for $U(\theta_1, \theta_2)$, $\mu = \frac{(\theta_1+\theta_2)}{2}$ and $\sigma^2 = \frac{(\theta_2-\theta_1)^2}{12}$

Proof:

Let y be the distance from the mid-point, $y = x - (\theta_2 + \theta_1)/2$, and the width be $w = \theta_2 - \theta_1$. Then since means add

$$\mu = \langle x \rangle = \frac{\theta_2 + \theta_1}{2} + \langle y \rangle = \frac{\theta_2 + \theta_1}{2} + \int_{-\frac{w}{2}}^{\frac{w}{2}} y \frac{1}{w} dy = \frac{\theta_1 + \theta_2}{2}.$$

Unsurprisingly the mean is the midpoint.

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_{-\frac{w}{2}}^{\frac{w}{2}} y^2 \frac{1}{w}\, dy = \frac{1}{w}\left[\frac{y^3}{3}\right]_{-\frac{w}{2}}^{\frac{w}{2}} = \frac{1}{3w}\left(\frac{w^3}{8} + \frac{w^3}{8}\right) = \frac{w^2}{12}$$
$$= \frac{(\theta_2 - \theta_1)^2}{12}$$

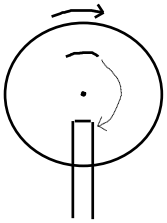
Occurrence of the Uniform distribution

1) Waiting times from random arrival time until a regular event (see below)

2) Engineering tolerances: e.g. if a diameter is quoted "±0.1mm", it sometimes assumed (probably incorrectly) that the error has a U(-0.1, 0.1) distribution.

3) Simulation: programming languages often have a standard routine for simulating the U(0, 1) distribution. This can be used to simulate other probability distributions.

Example: Disk wait times



In a hard disk drive, the disk rotates at 7200rpm. The wait time is defined as the time between the read/write head moving into position and the beginning of the required information appearing under the head.

(a) Find the distribution of the wait time.

(b) Find the mean and standard deviation of the wait time.

(c) Booting a computer requires that 2000 pieces of information are read from random positions. What is the total expected contribution of the wait time to the boot time, and rms deviation?

Solution

Rotation time = 8.33ms. Wait time can be anything between 0 and 8.33ms and each time in this range is as likely as any other time. Therefore, distribution of the wait time is U(0, 8.33ms) (i. . θ_1 = 0 and θ_2 = 8.33ms).

$$\mu = \frac{0+8.33}{2}\ \text{ms} = 4.2\ \text{ms}$$

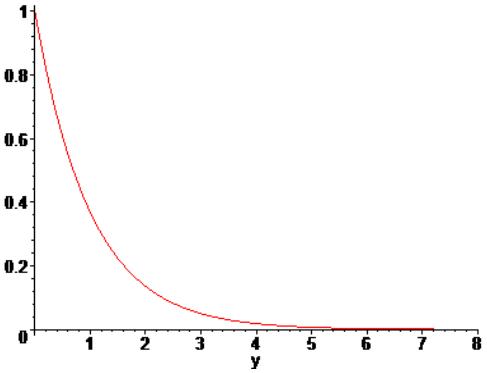
$$\sigma^2 = \frac{(8.33-0)^2}{12}\ \text{ms} = 5.8\ \text{ms}^2;\ \Rightarrow \sigma = 2.4\ \text{ms}$$

For 2000 reads the mean time is 2000 ×4.2 ms = 8.3s.
The variance is 2000 ×5.7ms^2= 0.012s^2, so σ =0.11s.

Exponential distribution

The continuous random variable Y has the Exponential distribution, parameter ν if:

$$f(y) = \begin{cases} \nu e^{-\nu y}, & y > 0 \\ 0, & y < 0 \end{cases}$$



Relation to Poisson distribution: If a Poisson process has constant rate ν, the mean after a time t is $\lambda = \nu t$. The probability of no-occurrences in this time is

$$P(k = 0) = e^{-\lambda} = e^{-\nu t}.$$

If $f(t)$ is the pdf for the first occurrence, then the probability of no occurrences is also given by

$$P(k=0) = 1 - P(\text{first occurrence has happened by } t) = 1 - \int_0^t f(t)dt$$

So equating the two ways of calculating the probability we have

$$e^{-\nu t} = 1 - \int_0^t f(t)dt$$

Now we can differentiate with respect to t giving

$$-\nu e^{-\nu t} = -\frac{d}{dt}\int_0^t f(t)dt = -f(t)$$

hence $f(t) = \nu e^{-\nu t}$: the time until the first occurrence (and between subsequent occurrences) has the Exponential distribution, parameter ν.

Occurrence

1) Time until the failure of a part.

2) Times between randomly happening events

Mean and variance

$$\mu = \int_{-\infty}^{\infty} y\, f(y)dy = \int_0^{\infty} y\nu e^{-\nu y}dy = [-ye^{-\nu y}]_0^{\infty} + \int_0^{\infty} e^{-\nu y}dy = \left[-\frac{e^{-\nu y}}{\nu}\right]_0^{\infty} = \frac{1}{\nu}$$

$$\sigma^2 = \int_{-\infty}^{\infty} y^2 f(y)dy - \mu^2 = \int_0^{\infty} y^2\nu e^{-\nu y}dy - \frac{1}{\nu^2}$$
$$= [-y^2e^{-\nu y}]_0^{\infty} + 2\int_0^{\infty} y\, e^{-\nu y}dy - \frac{1}{\nu^2} = 0 + 2\frac{\mu}{\nu} - \frac{1}{\nu^2} = \frac{1}{\nu^2}$$

Example: Reliability

The time till failure of an electronic component has an Exponential distribution and it is known that 10% of components have failed by 1000 hours.

(a) What is the probability that a component is still working after 5000 hours?

(b) Find the mean and standard deviation of the time till failure.

Solution

(a) Let *Y* = time till failure in hours; $f(y) = \nu e^{-\nu y}$

$$P(Y \leq 1000) = \int_0^{1000} \nu e^{-\nu y} = [-e^{-\nu y}]_0^{1000} = 1 - e^{-1000\nu} = 0.1$$

$$\Rightarrow e^{-1000\nu} = 0.9 \Rightarrow -1000\nu = \ln 0.9 = -0.10536$$

$$\Rightarrow \nu \approx 1.05 \times 10^{-4}$$

$$P(Y > 5000) = \int_{5000}^{\infty} \nu e^{-\nu y} dy = [-e^{-\nu y}]_{5000}^{\infty} = e^{-5000\nu} \approx 0.59$$

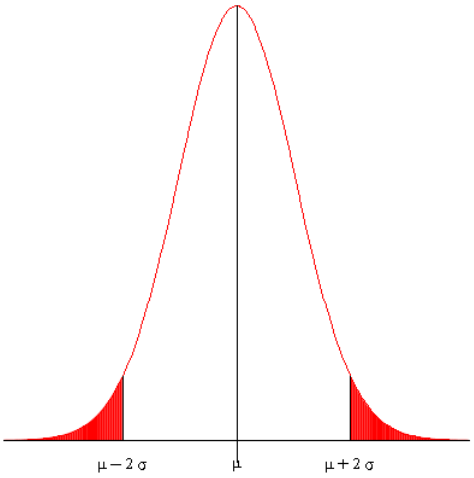
(b) Mean = $1/\nu$ = 9491 hours.

Standard deviation = $\sqrt{\text{Variance}}$ = $\sqrt{\frac{1}{\nu^2}} = 1/\nu$ = 9491 hours.

Normal distribution

The continuous random variable *X* has the Normal distribution if the pdf is:

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



The parameter μ is the mean and and the variance is σ^2. The distribution is also sometimes called a Gaussian distribution.

The pdf is symmetric about μ. *X* lies between μ - 1.96σ and $\mu + 1.96\sigma$ with probability 0.95 i.e. *X* lies within 2 standard deviations of the mean approximately 95% of the time.

Normalization
[*non-examinable*]

$f(x)$ cannot be integrated analytically for general ranges, but the full range can be integated as follows. Define

$$I = \int_{-\infty}^{\infty} dx \, \mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}} = \int_{-\infty}^{\infty} dx \, \mathrm{e}^{\frac{-x^2}{2\sigma^2}}$$

Then switching to polar co-ordinates we have

$$I^2 = \int_{-\infty}^{\infty} dx\, \mathrm{e}^{\frac{-x^2}{2\sigma^2}} \int_{-\infty}^{\infty} dy\, \mathrm{e}^{\frac{-y^2}{2\sigma^2}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy\, \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}} = \int_0^{\infty}\int_0^{2\pi} rdrd\theta\, \mathrm{e}^{-\frac{r^2}{2\sigma^2}} = 2\pi\int_0^{\infty} rdr\mathrm{e}^{-\frac{r^2}{2\sigma^2}}$$

$$= 2\pi\left[-\sigma^2 \mathrm{e}^{-\frac{r^2}{2\sigma^2}}\right]_0^{\infty} = 2\pi\sigma^2$$

Hence I = $\sqrt{2\pi\sigma^2}$ and the normal distribution integrates to one.

Mean and variance

The mean is μ because the distribution is symmetric about μ (or you can check explicitly by integrating by parts). The variance can be also be checked by integrating by parts:

$$\int_{-\infty}^{\infty}(x-\mu)^2 \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}dx = \int_{-\infty}^{\infty} y^2 \frac{e^{\frac{-y^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}dy = \int_{-\infty}^{\infty} y \times y \frac{e^{\frac{-y^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}dy$$

$$= \left[-y\sigma^2 \frac{e^{\frac{-y^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\sigma^2 \frac{e^{\frac{-y^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}dy = \sigma^2\int_{-\infty}^{\infty} f(y)dy = \sigma^2$$

Occurrence of the Normal distribution

1) Quite a few variables, e.g. human height, measurement errors, detector noise. (Bell-shaped histogram).

2) Sample means and totals - see below, Central Limit Theorem.

3) Approximation to several other distributions - see below.

Change of variable

The probability for X in a range dx around x is for a distribution $f(x)$ is given by $f(x)dx$. The probability should be the same if it is written in terms of another variable $y = y(x)$. Hence

$$f(x)dx = f(y)dy \Rightarrow f(y) = f(x)\frac{dx}{dy}.$$

Standard Normal distribution

There is no simple formula for $\int_a^b f(x)dx$, so numerical integration (or tables) must be used. The following result means that it is only necessary to have tables for one value of μ and σ^2.

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

This follows when changing variables since $\frac{dz}{dx} = \frac{1}{\sigma}$ hence

$$f(z) = f(x)\frac{dx}{dz} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \times \sigma = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}.$$

Z is the *standardised* value of *X*; N(0, 1) is the *standard Normal distribution*. The Normal tables give values of Q=P($Z \le z$), also called $\Phi(z)$, for *z* between 0 and 3.59.

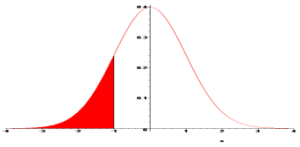
$$\mathrm{Q(z)} = \Phi(z) = P(Z \le z) = \int_{-\infty}^{z} f(x)dx$$

Outside of exams this is probably best evaluated using a computer package (e.g. Maple, Mathematica, Matlab, Excel); for historical reasons you still have to use tables.

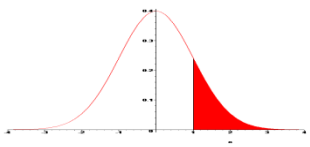
Example: Using standard Normal tables (on course web page and in exams)

If $Z \sim$ N(0, 1):

(a) $P(Z \le 1.0) = \Phi(1.0) = 0.8413$

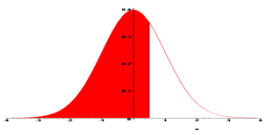


(b) $P(Z \le= -1.0) = P(Z \ge 1.0)$ (by symmetry)
$= 1 - P(Z < 1.0)$
= 1 - 0.8413 = 0.1587



(c) $P(Z > -0.5) = P(Z \le 0.5) = \Phi(0.5)$

= 0.6915.

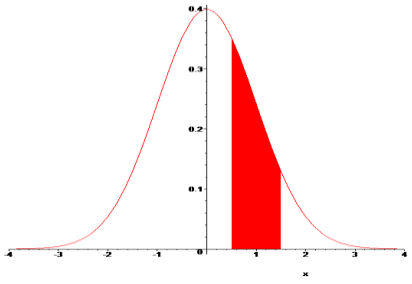


(d) $P(0.5 < Z < 1.5) = P(Z < 1.5) - P(Z < 0.5)$

= $\Phi(1.5)$ - $\Phi(0.5)$

= 0.9332 - 0.6915

= 0.2417.



(e) $P(Z < 1.356)$ - Using interpolation:

$$P(Z < 1.356) = P(Z < 1.35) + \frac{1.356-1.35}{1.36-1.35}\big(P(Z < 1.36) - P(Z < 1.35)\big) = 0.9125$$

(f) $0.8 = P(Z \leq c) = \Phi(c)$

Using tables "in reverse", $c \approx 0.842$.

(g) Finding a range of values within which Z lies with probability 0.95:

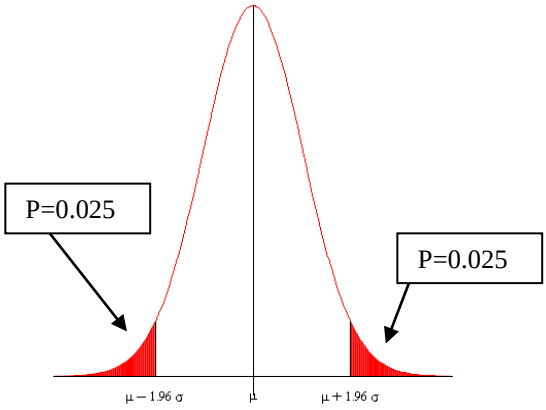
The answer is not unique; but suppose we want an interval which is symmetric about zero i.e. between $-d$ and d.

Tail area = 0.025

$\therefore$ P($Z \leq d$) = $\Phi(d)$ = 0.975

Using the tables "in reverse", d = 1.96.

$\therefore$ range is -1.96 to 1.96.



Example: Manufacturing variability

The outside diameter, X mm, of a copper pipe is N(15.00, 0.02^2) and the fittings for joining the pipe have inside diameter Y mm, where $Y \sim$ N(15.07, 0.022^2).

(i) Find the probability that X exceeds 14.99 mm.

(ii) Within what range will X lie with probability 0.95?

(iii) Find the probability that a randomly chosen pipe fits into a randomly chosen fitting (i.e. $X < Y$).

Solution

(i) $X \sim N(15.0, 0.02^2)$

$$P(X > 14.99) = P\left(Z > \frac{14.99 - 15.0}{0.02}\right) = P(Z > -0.5) = P(Z < 0.5) \approx 0.6915$$

(ii) From previous example $Z = \frac{X-15.0}{0.02}$ lies in (-1.96, 1.96) with probability 0.95.
i.e. $P\left(-1.96 < \frac{X-15.0}{0.02} < 1.96\right) = 0.95$

$$\Rightarrow P(15 - 0.02 \times 1.96 < X < 15 + 0.02 \times 1.96) = 0.95$$

$$\Rightarrow P(14.96 < X < 15.04) = 0.95$$

i.e. the required range is 14.96mm to 15.04mm.

(iii) For $X < Y$ we want $P(Y - X > 0)$. To answer this we need to know the distribution of $Y - X$.

Distribution of the sum of Normal variates

Remember than means and variances of independent random variables just add. So if $X_1, X_2, \ldots, X_n$ are independent and each have a normal distribution $X_i \sim N(\mu_i, , \sigma_i^2)$, we can easily calculate the mean and variance of the sum. A special property of the Normal distribution is that the distribution of the sum of Normal variates is *also* a Normal distribution. So if $c_1, c_2, \ldots, c_n$ are constants then:

$$c_1X_1 + c_2X_2 + \ldots + c_nX_n \sim \mathrm{N}(c_1\mu_1 + \ldots + c_n\mu_n, c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots c_n^2\sigma_n^2)$$

Proof that the distribution of the sum is Normal is beyond scope. Useful special cases for two variables are

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$
$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

If all the X's have the same distribution i.e. $\mu_1 = \mu_2 = \ldots = \mu_n = \mu$, say and $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$, say, then:

(iii) All $c_i = 1$: $X_1 + X_2 + \ldots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$

(iv) All $c_i = 1/n$: $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n} \sim \mathrm{N}(\mu, \sigma^2/n)$

The last result tells you that if you average n identical independent noisy measurements, the error decreases by $1/\sqrt{n}$. (variance goes down as $1/n$).

Example: Manufacturing variability (iii)

Find the probability that a randomly chosen pipe fits into a randomly chosen fitting (i.e. $X < Y$).

Using the above results

$$Y - X \sim N(\mu_Y - \mu_X, \sigma_Y^2 + \sigma_X^2) = N(15.07 - 15, 0.02^2 + 0.022^2)$$
$$= N(0.07, 0.000884)$$

Hence

$$P(Y - X > 0) = P\left(Z > \frac{0 - 0.07}{\sqrt{0.0.000884}}\right) = P(Z > -2.35) = P(Z < 2.35) \approx 0.991$$

Example: detector noise

A detector on a satellite can measure T+g, the temperature T of a source with a random noise g, where g ~ N(0, $1K^2$). How many detectors with independent noise would you need to measure T to an rms error of 0.1K?

Answer: We can estimate the temperature from *n* detectors by calculating the mean from each. The variance of the mean will be $1K^2/n$ where *n* is the number of detectors. An rms error of 0.1K corresponds to a variance of 0.01 K^2, hence we need *n*=100 detectors.

Normal approximations

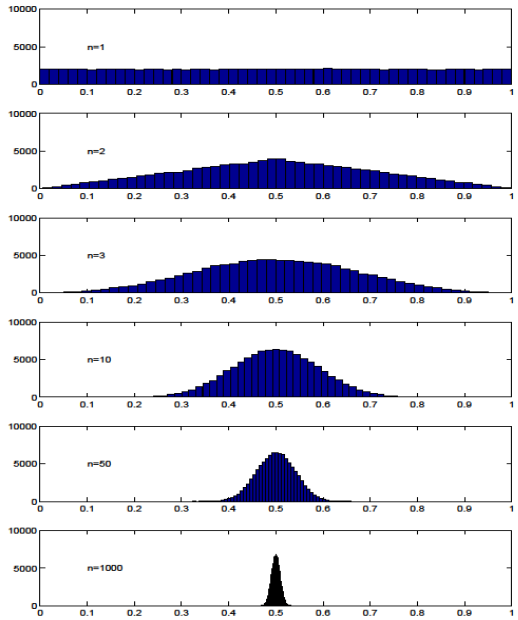
Central Limit Theorem: If X_1, X_2, ... are independent random variables with the same distribution, which has mean μ and variance σ^2 (both finite), then the sum $\sum_{i=1}^{n} X_i$ tends to the distribution $N(n\mu, n\sigma^2)$ as $n \to \infty$.

Hence: The sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is distributed approximately as N(μ, σ^2/n) for large *n*.

For the approximation to be good, *n* has to be bigger than 30 or more for skewed distributions, but can be quite small for simple symmetric distributions.

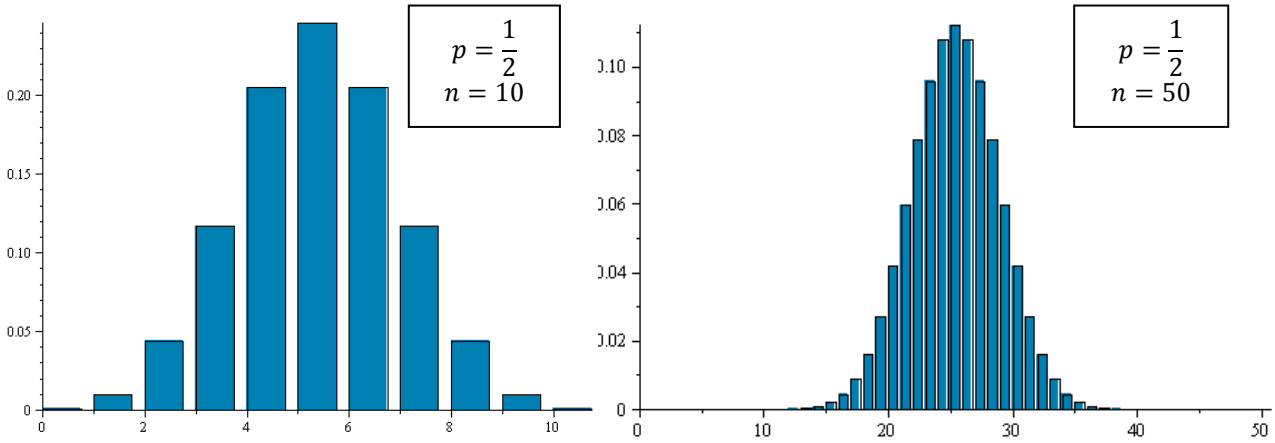
The approximation tends to have much better fractional accuracy near the peak than in the tails: don't rely on the approximation to estimate the probability of very rare events.

Example: Average of *n* samples from a uniform distribution:

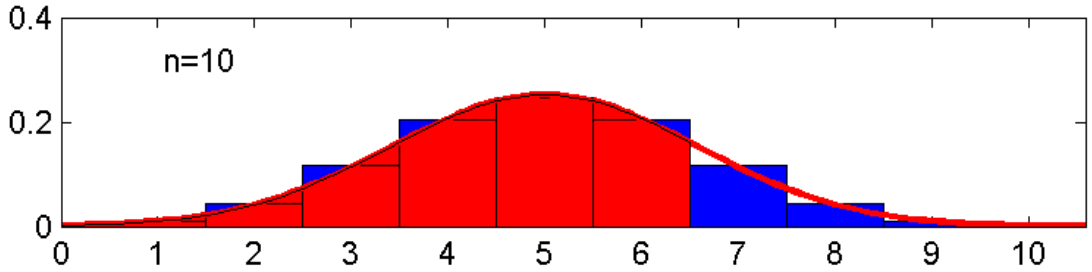


Normal approximation to the Binomial

If $X \sim B(n, p)$ and n is large and np is not too near 0 or 1, then X is approximately $N(np, np(1-p))$.



The probability of getting k from the Binomial distribution can be approximated as the probability under a Normal distribution for getting x in the range from $k - \frac{1}{2}$ to $k + \frac{1}{2}$. For example $P(k \leq 6)$ can be approximated as $\int_{-\infty}^{6.5} f(x)dx$ where $f(x)$ is the Normal distribution:



Example: I toss a coin 1000 times, what is the probability that I get more than 550 heads?

Answer: The number of heads has a binomial distribution with mean np=500 and variance $np(1-p) = 250$ So the number of heads H can be approximated as $H \sim N(500,250)$. Hence

$$P(H > 550) \approx P\left(Z > \frac{550.5 - 500}{\sqrt{250}}\right) \approx P(Z > 3.19) \approx 1 - 0.9993 \approx 0.007.$$

Quality control example:

The manufacturing of computer chips produces 10% defective chips. 200 chips are randomly selected from a large production batch. What is the probability that fewer than 15 are defective?

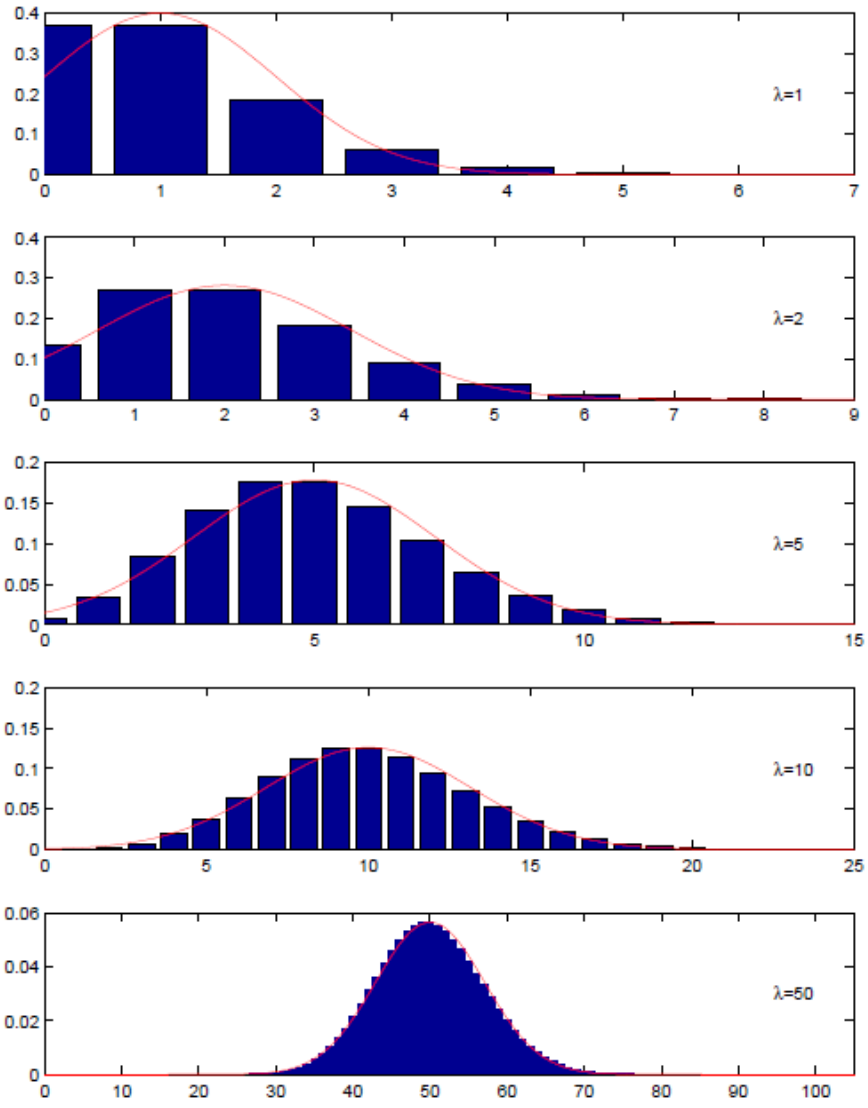
Answer: the mean is $np = 200 \times 0.1 = 20$, variance $np(1-p) = 200 \times 0.1 \times 0.9 = 18$. So if X is the number of defective chips, approximately $X \sim N(20,18)$, hence

$$P(X < 15) \approx P\left(Z < \frac{14.5 - 20}{\sqrt{18}}\right) = P(Z < -1.296) = 1 - P(Z < 1.296)$$
$$= 1 - [0.9015 + 0.6 \times (0.9032 - 0.9015)] \approx 0.097$$

This compares to the exact Binomial answer $\sum_{k=0}^{14} C_k^n p^k (1-p)^{n-k} \approx 0.093$. The Binomial answer is easy to calculate on a computer, but the Normal approximation is *much* easier if you have to do it by hand. The Normal approximation is about right, but not accurate.

Normal approximation to the Poisson

If $Y \sim$ Poisson parameter λ and λ is large (> 7, say), then Y has approximately a $N(\lambda, \lambda)$ distribution.



Example: Stock Control

At a given hospital, patients with a particular virus arrive at an average rate of once every five days. Pills to treat the virus (one per patient) have to be ordered every 100 days. You are currently out of pills; how many should you order if the probability of running out is to be less than 0.005?

Solution

Assume the patients arrive independently, so this is a Poisson process, with rate 0.2 / day.

Therefore, *Y*, number of pills needed in 100 days, ~ Poisson, λ = 100 x 0.2 = 20.

We want $P(Y > n) < 0.005$, or $P\left(Y \leq n + \frac{1}{2}\right) > 0.995$ under the Normal approximation, where a probability of 0.995 corresponds (from tables) to z =2.575. Since $\mu = \sigma = \lambda = 20$ this corresponds to. $n + \frac{1}{2} = 20 + 2.575\sqrt{20} = 31.5$, so we need to order $n \geq 32$pills.

Comment

Let's say the virus is deadly, so we want to make sure the probability is less than 1 in a million, 10^{-6}. A normal approximation would give 4.7σ above the mean, so $n \geq 42$ pills. But surely getting just a bit above twice the average number of cases is not that unlikely??

Yes indeed, the assumption of independence is extremely unlikely to be valid. Viruses tend to be infectious, so occurrences are definitely not independent. There is likely to be a small but significant probability of a large number of people being infected simultaneously – a much larger number of pills needs to be stocked to be safe.

Don't use approximations that are too simple if their failure might be important! Rare events in particular are often a lot more likely than predicted by (too-) simple approximations for the probability distribution.