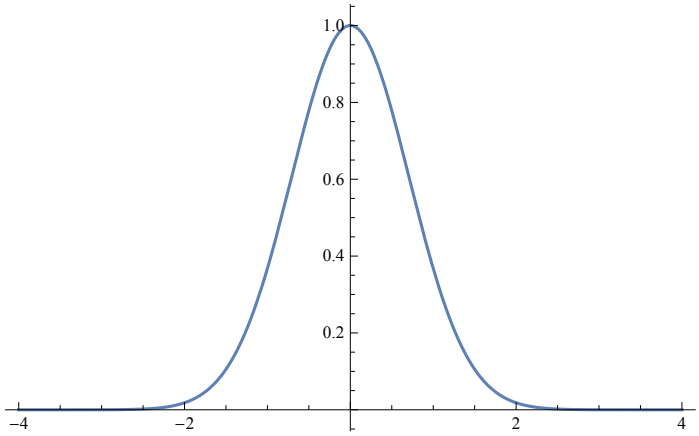
11 The normal distribution and the central limit theorem

11.1 The Normal Distribution

The density of the normal distribution is related to the function e^{-x^2}. Its graph looks like the following.

```
Plot[e^-x^2, {x, -4, 4}]
```



The bad thing about e^{-x^2} is that its integral is not an elementary function. The integral can be expressed in terms of the error function, *erf*(*x*), which is defined by $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$. One has $\int e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2} erf(x)$.

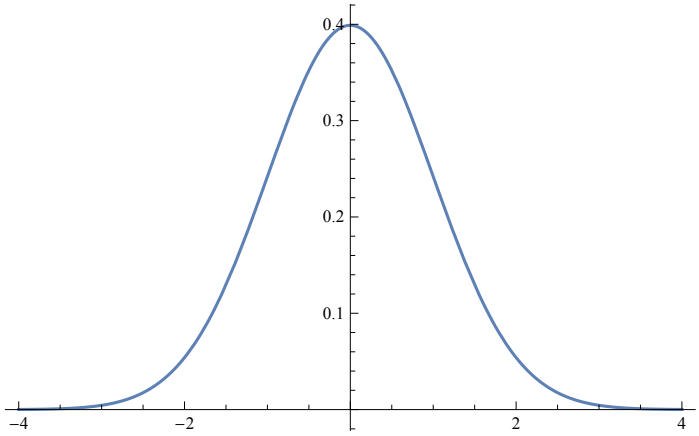
```
Integrate[ E^(-x^2), x]
```

$$\frac{1}{2}\sqrt{\pi}\ \text{Erf[x]}$$

Even though the integral of e^{-x^2} can not be expressed in terms of elementary functions, it is possible to show that $\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$. Therefore $\frac{1}{\sqrt{\pi}} e^{-x^2}$ is a probability density function. It has mean 0 because its graph is symmetrical about x = 0. The variance is $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2\, e^{-x^2}\, dx$. If we integrate by parts letting u = x and dv = xe^{-x^2} we see that this integral is equal to $\frac{1}{\sqrt{\pi}}$ [$\frac{-xe^{-x^2}}{2}\Big|_{-\infty}^{\infty}$ + $\frac{1}{2}\int_{-\infty}^{\infty} e^{-x^2}\, dx$] = $\frac{1}{2}$. Thus the variance is equal to 1/2 and the standard deviation is equal to $1/\sqrt{2}$. We can get a density function whose standard deviation is 1 by replacing x by $x/\sqrt{2}$ and multiplying by an appropriate factor so its integral is 1. The appropriate factor is $1/\sqrt{2}$. This gives us the standard normal density function f(x) = $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

```
f[x_] = e^(-x^2/2) / Sqrt[2 π]
Plot[f[x], {x, -4, 4}]
∫_-2^2. f[x] dx
```

$$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\,\pi}}$$




```
0.9545
```

As you can see. $\frac{1}{\sqrt{2\pi}}\int_{-2}^{2} e^{-x^2/2}\,dx = 0.9545$. So about 95% of the time an observation from the standard normal distribution will be within two standard deviations from the mean.

We can get a density function whose mean is μ and standard deviation is σ by replacing x by (x - μ)/σ and multiplying by an appropriate factor so its integral is 1. The appropriate factor is $1/\sigma$. This gives us the normal density function with mean μ and standard deviation σ $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-(x-\mu)^2/2\,\sigma^2}$.

```
f[x_, μ_, σ_] = (1 / (Sqrt[2 Pi] * σ)) E^(-((x - μ)^2) / (2 σ^2))
```

$$\frac{e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}}{\sqrt{2\,\pi}\ \sigma}$$

It turns out that again about 95% of the time an observation from a normal distribution will be within two standard deviations from the mean. We showed this above for the standard normal distribution. The general case can be reduced to te standard case by a change of variables in the integral.

Example 7.4 on page 229 of Meerschaert: Suppose the times between fires in Example 1 of section 2 above are modeled by a normal distribution with mean 4 and standard deviation 4, so that the density function is $f(t;\,4,\,4) = \frac{1}{\sqrt{2\pi}\ 4} e^{-(t-4)^2/2\,(4)^2}$. What is the probability that the time between two fires will be between 3 and 5?

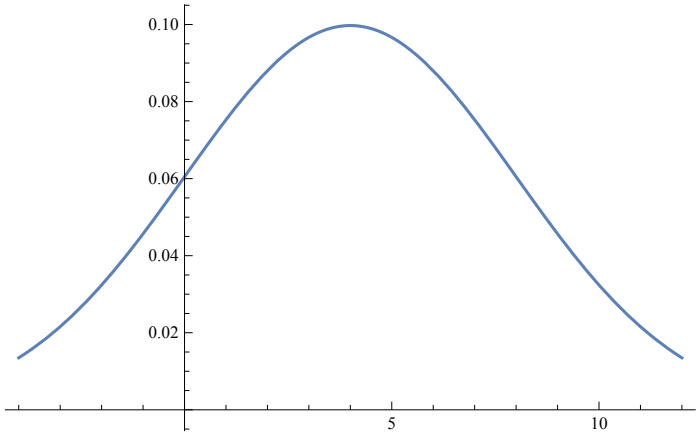
Solution: $\int_3^5 \frac{1}{\sqrt{2\pi}\ 4} e^{-(t-4)^2/2\,(4)^2}\,dt$. It we let y = (t - 4)/4, then the integral becomes $\int_{-1/4}^{1/4} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy = 2\int_0^{1/4} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy$.

This can either be evaluated using the error function or one can use a table of the cummulative distribution of the standard normal distribution which can be found in most statistics books. In terms of the cummulative distribution F(x) the above integral is 2[F(1/4) - 0.5] = 2[0.5987 - 0.5] = 2 [0.0987] = 0.1974. I used a table in a statistics book to get F(1/4). Here is the same calculation

using *Mathematica*.

```
f[t, 4, 4]
Plot[f[t, 4, 4], {t, -4, 12}]
∫_3^5 f[t, 4.`, 4] dt
```

$$\frac{e^{-\frac{1}{32}(-4+t)^2}}{4\sqrt{2\pi}}$$




```
0.197413
```

Here is another way using *Mathematica*.

```
<< Statistics`ContinuousDistributions`
f[t_] = PDF[NormalDistribution[4, 4], t]
g[t_] = CDF[NormalDistribution[4, 4], t]
g[5.] - g[3]
```

```
Get::noopen : Cannot open Statistics`ContinuousDistributions`. >>

$Failed
```

$$\frac{e^{-\frac{1}{32}(-4+t)^2}}{4\sqrt{2\pi}}$$

$$\frac{1}{2}\,\text{Erfc}\left[\frac{4-t}{4\sqrt{2}}\right]$$

```
0.197413
```

11.2 The Central Limit Theorem

Suppose $X_1, X_2, \ldots, X_n, \ldots$ is an infinite sequence of independent random variables each having the same probability mass or density function f(x). Suppose we form their "averages" $\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$. The law of large numbers says that $\overline{X}_n \to \mu$ with probability 1. The central limit theorem says that the distribution of $\overline{X}_n$ is approximately normal with the same mean and standard deviation $\sigma/\sqrt{n}$. A more precise statement is that $\Pr\{ \mu + a/\sqrt{n} < \overline{X}_n < \mu + b/\sqrt{n} \} \to \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/2\sigma^2}\, dx$ as $n \to \infty$.

Example: An emergency 911 service in a local community received an average of 171 calls per month for house fires over the past

year. On the basis of this data, the rate of house fire emergencies was estimated at 171 per month. The next month only 153 calls were received. Does this indicate an actual reduction in the rate of house fires, or is it simply a random fluctuation?

Solution: Suppose the time between house fires is an exponential random variable with mean λ = 171fires / month. So the mean time between fires is $\mu = \frac{1}{\lambda} = \frac{1}{171} \approx 0.00585$ months and σ is also $\frac{1}{\lambda} = \frac{1}{171} \approx 0.00585$ months. Now in the next month the mean time $\bar{t}$ between fires in a sample of 153 fires is $\frac{1}{153} \approx 0.00654$. According to the central limit theorem the mean of 153 observations should be approximately normally distributed with mean $\mu = 0.00585$ and $\sigma = \frac{0.00585}{\sqrt{153}} \approx 0.000473$.

```
Print[1/171., "    ", 1/153., "    ", 1/(171. * Sqrt[153])]
```

```
0.00584795    0.00653595    0.000472779
```

As noted above, about 95% of the time an observation from a normal distribution will be within two standard deviations from the mean, i.e in the interval $\mu - 2\sigma \le x \le \mu + 2\sigma$. This is called the 95% confidence interval. Only about 5% of the time will it be outside this interval. So a standard test that an observation does not come from a certain normal distribution is that it is over two standard deviations from the mean.

In our example the 95% confidence interval is 0.00585 - 2 * 0.000473 ≤ t ≤ 0.00585 + 2 * 0.000473 which is

```
Print[1/171. - 2/(171. * Sqrt[153]), " ≤ t ≤ ", 1/171. + 2/(171. * Sqrt[153])]
```

```
0.0049024 ≤ t ≤ 0.00679351
```

In our example the observation was 0.00654. This is in the 95% confidence interval so we attribute this to random fluctuation.

Problem 1. Customers arrive at a bank on the average of one every three minutes. Assume arrivals are independent of each other and have exponential distributions. Use the central limit theorem to find approximately the probability that between 75 and 85 customers arrive in the next four hours. Assume 75 and 85 are included in "between 75 and 85".