Estimating binomial proportions

Roberto Rossi
University of Edinburgh Business School
Edinburgh, UK

Problem description

Consider the problem of estimating the parameter p of a random variable which follows a Binomial distribution Bin(M, p), where M is known. The estimation should be carried out by exploiting information from K past observations of the random variable.

- **Example**

```
M = 100;
p = 0.5;
K = 10;
SeedRandom[1234];
observations = RandomInteger[BinomialDistribution[M, p], K];
Print["Observations: " <> ToString[observations]];
```

```
Observations: {40, 48, 42, 48, 47, 45, 52, 57, 42, 55}
```

In the given example, we aim to estimate parameter p, by using the $K = 10$ past observations for the Binomial random variable.

Maximum likelihood estimation

A common approach to carry out this estimation consists in determining the "maximum likelihood estimator" $\hat{p}$ for p. In the case of Binomial proportion, this is known to be simply

$$\hat{p} = \frac{\sum_{i=1}^{K} x_i}{M\,K}$$

where x_i denotes the i-th observation for the random variable. In our previous example, the maximum likelihood estimator for p is

```
Print["Maximum likelihood: " <> ToString[N[Total[observations] / (M * K)]]];
```

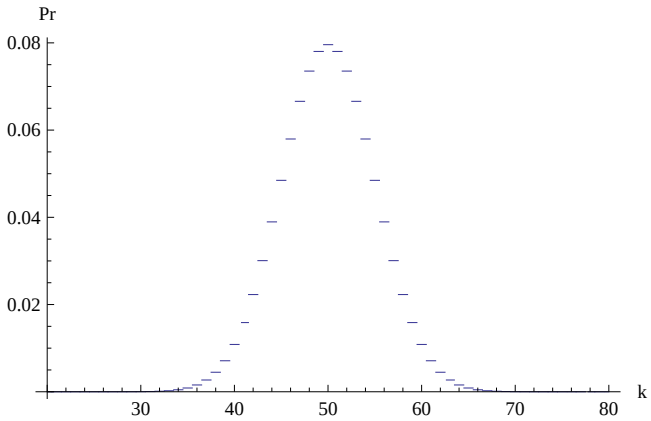
```
Maximum likelihood: 0.476
```

This approach is based on the concept of **likelihood function**. To explain this concept, first consider the **probability mass function** of our Binomial random variable r, this is simply

$$\Pr\{r = k\} = \binom{M}{k} p^k (1-p)^{M-k}.$$

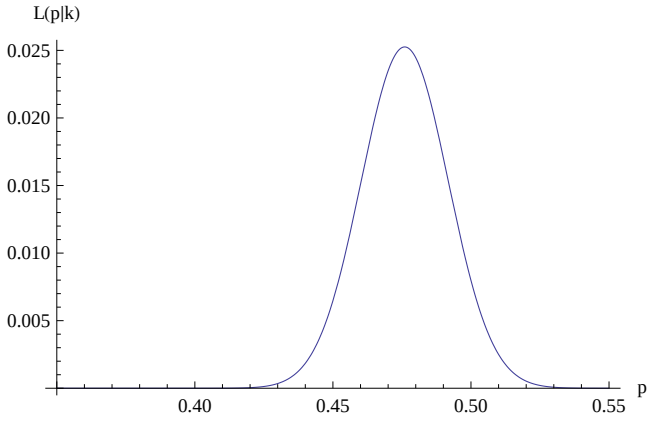
We can plot this function for our previous example.

```
Plot[PDF[BinomialDistribution[M, p], Round[k]], {k, 20, 80}, AxesLabel → {"k", Pr}]
```



In the previous plot, we varied the number of successes k and for each possible value in the support of r we plotted the respective probability. If, conversely, we fix k and we vary p what we obtain is the so-called, likelihood function. We denote this function as $L(p \mid k)$. It should be noted that the likelihood function **does not** represent a probability density function for p. We now plot the likelihood function for $k = \sum_{i=1}^{K} x_i$.

```
k = Total[observations];
Plot[{PDF[BinomialDistribution[M * K, p], k]}, {p, 0.35, 0.55}, AxesLabel → {"p", "L(p|k)"}]
```



Intuitively, the maximum likelihood estimator is the value of p which maximizes this function. Therefore,

```
Print["Maximum likelihood: " <>
  ToString[NArgMax[PDF[BinomialDistribution[M * K, i], k], {i, 0, 1}][[1]]]]
```

```
Maximum likelihood: 0.476
```

This value corresponds to the value we have previously computed in closed form. We now prove that, indeed, the closed form expression of this estimator is

$$\hat{p} = \frac{\sum_{i=1}^{K} x_i}{M\,K}.$$

■ Proof

$$L\left(p \mid \sum_{i=1}^{K} x_i\right) = \prod_{i=1}^{K} \Pr\{r = x_i\}$$

$$L\left(p \mid \sum_{i=1}^{K} x_i\right) = \prod_{i=1}^{K} \left(p^{x_i}(1-p)^{M-x_i}\right)$$

$$L\left(p \mid \sum_{i=1}^{K} x_i\right) = p^{\sum_{i=1}^{K} x_i}(1-p)^{M\,K-\sum_{i=1}^{K} x_i}$$

$$\ln L\left(p \mid \sum_{i=1}^{K} x_i\right) = \sum_{i=1}^{K} x_i \ln p + \left(M\,K - \sum_{i=1}^{K} x_i\right) \ln(1-p)$$

$$\frac{d}{dp} \ln L\left(p \mid \sum_{i=1}^{K} x_i\right) = \frac{1}{p} \sum_{i=1}^{K} x_i - \frac{1}{1-p}\left(M\,K - \sum_{i=1}^{K} x_i\right) = 0$$

$$\frac{d}{dp} \ln L\left(p \mid \sum_{i=1}^{K} x_i\right) = \frac{(1-p)\sum_{i=1}^{K} x_i - p\left(M\,K - \sum_{i=1}^{K} x_i\right)}{p(1-p)} = 0$$

$$\frac{d}{dp} \ln L\left(p \mid \sum_{i=1}^{K} x_i\right) = \sum_{i=1}^{K} x_i - p\sum_{i=1}^{K} x_i - p\,M\,K + p\sum_{i=1}^{K} x_i = 0$$

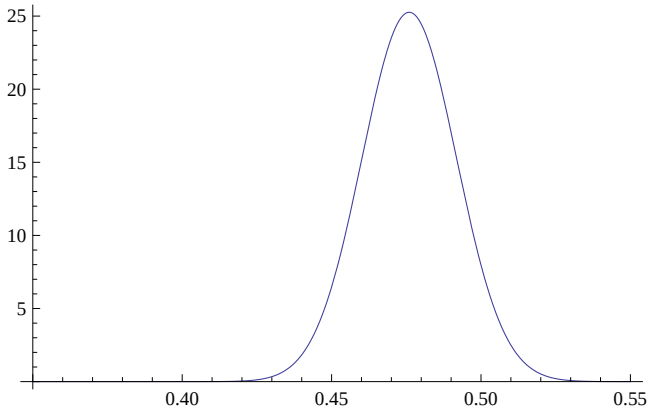
$$\frac{d}{dp} \ln L\left(p \mid \sum_{i=1}^{K} x_i\right) = \sum_{i=1}^{K} x_i - p\,M\,K = 0$$

$$\hat{p} = \frac{\sum_{i=1}^{K} x_i}{M\,K} \quad \square$$

■ Normal approximation

Because of the central limit theorem, it is clearly possible to use a Normal distribution in place of the Binomial distribution to carry out the computations previously discussed. In particular, the approximate likelihood function $\tilde{L}(p \mid k)$ will be the likelihood function of a Normal distribution with mean p and standard deviation $\sqrt{(p(1-p))/(M\,K+1)}$ computed for $k = \frac{\sum_{i=1}^{K} x_i}{M\,K}$. We plot this function and compute the value of p for which a maximum is reached.

```
Plot[{PDF[NormalDistribution[p, Sqrt[p * (1 - p)/(M K)]], Total[observations] / (M K)]},
 {p, 0.35, 0.55}]
Print["Maximum likelihood: " <> ToString[NArgMax[
    PDF[NormalDistribution[i, Sqrt[i * (1 - i)/(M K)]], Total[observations] / (M K)], {i, 0, 1}][[1]]]]
```




```
Maximum likelihood: 0.475976
```

This demonstrate the high quality of the Normal approximation.

Bayesian estimation

According to Jaynes' principle of maximum entroy, we shall assign to p a uniform prior distribution to express our uncertainty about its true value. A uniform prior is also known as "uninformative" as it is the maximum entropy distribution among the continuous ones with support over (0, 1). It should be noted that in Bayesian analysis the prior distribution expresses uncertainty about the true value of p and does not attribute randomness to p. The analysis then proceeds as follows: we multiply the prior distribution - which in this case is simply a constant equal to 1 since we consider a uniform prior for $p \in (0, 1)$ with pdf $(p) = 1$ - by the likelihood function and then we "normalize" to obtain the posterior distribution. Therefore recall that the likelihood function is

$$L\left(p \mid \sum_{i=1}^{K} x_i\right) = \prod_{i=1}^{K} \Pr\{r = x_i\} = \prod_{i=1}^{K} \left(p^{x_i}(1-p)^{M-x_i}\right) = p^{\sum_{i=1}^{K} x_i}(1-p)^{M K - \sum_{i=1}^{K} x_i}$$

If we integrate the likelihood function as follows

$$\int_0^1 p^{\sum_{i=1}^{K} x_i}(1-p)^{M K - \sum_{i=1}^{K} x_i}\, dp = \frac{\left(\sum_{i=1}^{K} x_i\right)!\left(M K - \sum_{i=1}^{K} x_i\right)!}{(M K+1)!}$$

we obtain an area generally different than one, therefore we must use this value to "normalize" the likelihood function in order to obtain the following posterior distribution

$$f\left(p \mid \sum_{i=1}^{K} x_i\right) = \frac{(M K+1)!}{\left(\sum_{i=1}^{K} x_i\right)!\left(M K - \sum_{i=1}^{K} x_i\right)!}\; L\left(p \mid \sum_{i=1}^{K} x_i\right) = \frac{(M K+1)!}{\left(\sum_{i=1}^{K} x_i\right)!\left(M K - \sum_{i=1}^{K} x_i\right)!}\; p^{\sum_{i=1}^{K} x_i}(1-p)^{M K - \sum_{i=1}^{K} x_i}$$

which is a Beta distribution with expected value

$$\int_0^1 p\, \frac{(M K+1)!}{\left(\sum_{i=1}^{K} x_i\right)!\left(M K - \sum_{i=1}^{K} x_i\right)!}\; p^{\sum_{i=1}^{K} x_i}(1-p)^{M K - \sum_{i=1}^{K} x_i}\, dp = \frac{\sum_{i=1}^{K} x_i + 1}{M K+2}.$$

It can be easily observed that the Bayesian analysis carried out is equivalent to Laplace's rule of succession.

Confidence interval analysis

Bayesian analysis falls short in the fact that it does not quantify the uncertainty associated with the most recent update for p. Assume you have 10 past observations of the random variable r, or 1000 past observations for it. It is fairly intuitive to grasp the fact that in the second case, you estimate will be far more accurate. Bayesian estimates simply do not capture this key aspect of estimation, since they are point estimates. No matter if you have 10 or 1000 past observations. What you get out of Laplace's rule of succession is a scalar number. In practice, statisticians never work with scalar numbers. They do consider a degree of confidence they want to achieve, and a confidence region with radius ϑ within which the true value of p is likely to lie according to the chosen confidence level α. Given α and a set of data, it is therefore possible to uniquely determine the confidence regision $\hat{p}\pm\vartheta$ within which the true value of p lies according to the prescribed confidence level α. The more data one has, the smaller this region will be. The fewer data one has, the larger this region will be. Furthermore, given a fixed set of data, increasing α will also enlarge the region (i.e. increase ϑ), while decreasing α will shrink the region (i.e. decrease ϑ). Modeling the uncertainty associated with a set of observation is an intrinsic two-dimensional matter. It is not possible, nor auspicable, to completely ignore the uncertainty associated with the data and come up with a point-wise estimates that gives no idea of the quality of the estimation carried out. For this reason, we now introduce a set of strategies for computing confidence intervals for Binomial proportions when a given set of data is available.

The first strategy for building confidence intervals for binomial proportions was proposed by Clopper and Pearson and operates as follows. Let r be a random variable distributed according to a Binomial$(M\ K,\ p)$. Consider the two values

$$p_{\text{ln}} = \min\left\{p \,\middle|\, \Pr\left\{r \geq \sum_{i=1}^{K} x_i\right\} \geq (1-\alpha)/2\right\}$$
$$p_{\text{ub}} = \max\left\{p \,\middle|\, \Pr\left\{r \leq \sum_{i=1}^{K} x_i\right\} \geq (1-\alpha)/2\right\}$$

Consider once more the example above, the α=0.95 Clopper and Pearson interval is

```
Print["plb: " <> ToString[InverseCDF[
      BetaDistribution[Total[observations], M K - Total[observations] + 1], (1 - 0.95) / 2]]];
Print["pub: " <> ToString[InverseCDF[BetaDistribution[Total[observations] + 1,
       M K - Total[observations]], (1 + 0.95) / 2]]];
```

```
plb: 0.444656
```

```
pub: 0.507486
```

The computation is executed by using the Beta distribution as a proxy, as discused by Agresti and Coull. These authors also discuss approximate, but very effective, strategies for computing confidence intervals for binomial proportions. The first strategy uses the following Normal approximation. Let $\hat{p}=\frac{\sum_{i=1}^{K} x_i}{M\ K}$ then the interval is computed as

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{M\ K}}$$

where $z_{1-\alpha/2}$ is the 1 - α/2 quantile of the standard normal distribution. Therefore

```
p̂ = Total[observations] / (M K);
Print["plb: " <> ToString[InverseCDF[NormalDistribution[p̂, Sqrt[p̂ (1 - p̂) / (M K)]], (1 - 0.95) / 2]]];
Print["pub: " <> ToString[InverseCDF[NormalDistribution[p̂, Sqrt[p̂ (1 - p̂) / (M K)]], (1 + 0.95) / 2]]];
```

```
plb: 0.445046

pub: 0.506954
```

Of course, the Normal approximation fails when the sequence observed comprises as many successes (or failures) as the number of trials. That is if $\frac{\sum_{i=1}^{K} x_i}{M\,K}$=1. Other approximate intervals which perform better in these limiting cases are Wilson's interval and Agresti-Coull interval. Agresti-Coull interval is a simple rule of thumb that prescribes to "add 2 successes and 2 failures" before computing the interval according to the Normal approximation strategy presented above. In practice, for our example this generates the following interval

```
p̂ = (Total[observations] + 2) / (M K + 4);

Print["plb: " <> ToString[InverseCDF[NormalDistribution[p̂, Sqrt[p̂ (1 - p̂) / (M K)]], (1 - 0.95) / 2]]];

Print["pub: " <> ToString[InverseCDF[NormalDistribution[p̂, Sqrt[p̂ (1 - p̂) / (M K)]], (1 + 0.95) / 2]]];
```

```
plb: 0.445141

pub: 0.50705
```

The interval becomes more conservative and displays better performances when the actual p is very close to 0 or 1.

Estimation when a long sequence of successes (failures) is observed

As previously discussed, the estimation becomes particularly complex when the original p we are trying to estimate is very close to 0 (a specular reasoning applies to p very close to 1). In such a case, it is possible to observe very long sequences of realizations equal to 0. A Normal approximation in this case fails badly. Also a maximum likelihood estimator does not perform well. The Bayesian strategy above is more conservative and let the estimated p value decay slowly, as new evidence is aquired. We will now discuss how the Bayesian strategy above simplifies in present of long sequences of zero observations. We will also discuss how the value of p decays and how we can use a Normal approximation strategy to handle this situation.

We shall now introduce an example in which the p we are trying to estimate is a very small value.

▪ Example

In[7]:=
```
M = 100;
p = 0.0000001;
K = 100;
SeedRandom[1234];
observations = RandomInteger[BinomialDistribution[M, p], K];
Print["Observations: " <> ToString[observations]];
```

```
Observations: {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
```

Once more, we aim to estimate parameter p, by using the $K = 100$ past observations for the Binomial random variable. Despite p being not 0, all we observed is a long sequence of zeroes. In such a situation, when a long sequence of zeroes is observed, the probability mass function simplifies as follows

$$\Pr\{r = 0\} = (1 - p)^M$$

The likelihood function also simplifies

$$L(p \mid 0) = \prod_{i=1}^{K} \Pr\{r = 0\} = \prod_{i=1}^{K} (1 - p)^M = (1 - p)^{M K}$$

By noting that the integral

$$\int_0^1 (1 - p)^{M K} \, dp = \frac{1}{M K + 1}$$

the posterior distribution for p, assuming a uniform prior distribution, is simply

$$f(p \mid 0) = \frac{\int_0^1 p(1-p)^{M K} \, dp}{\int_0^1 (1-p)^{M K} \, dp} = (M K + 1) \int_0^1 p(1 - p)^{M K} \, dp = \frac{1}{M K+2}$$

Therefore for our example the estimated value of p is

In[15]:=
```
NIntegrate[i * PDF[BinomialDistribution[Round[M K], i], 0], {i, 0, 1}] Round[M K + 1]
N[1 / (M K + 2)]
```

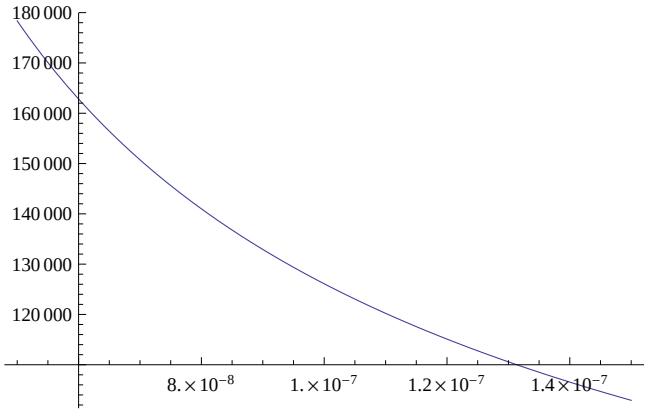
Out[15]=
```
0.00009998
```

Out[16]=
```
0.00009998
```

In contrast to a maximum likelihood estimator equal to 0. It should be noted that the Bayesian interpretation is not the only possible one. If we consider the likelihood function literally as a sort of "witness" for the likelihood of a particular value of p in (0,1), then what we can do is to normalize this function and use it to compute a weighted average on the possible values of p, values of p to which the witness function associates higher weights will be favoured. This interpretation for the above approach, despite giving the same result as the Bayesian approach, does not require to introduce a prior distribution and is only based on evidence gathered from the likelihood function.

We now adopt, as previously done, a Normal approximation to the likelihood function. We first consider the probability density function of a Normal distribution with mean p and standard deviation $\sqrt{(p(1 - p))/(M K + 1)}$ computed at 0. Then we plot this function for different values of p.

```
Plot[{PDF[NormalDistribution[p, Sqrt[p * (1 - p) / (M K)]], Total[observations] / (M K)]},
 {p, 0.00000005, 0.00000015}]
```



We then integrate this approximate likelihood function between 0 and 1 and normalize as usual. The resulting function is

$$\tilde{f}(p \mid 0) = \frac{\int_0^1 p \frac{e^{-\frac{p M K}{2(1-p)}}}{\sqrt{2\pi}\sqrt{\frac{(1-p)p}{MK+1}}} dp}{\int_0^1 \frac{e^{-\frac{p M K}{2(1-p)}}}{\sqrt{2\pi}\sqrt{\frac{(1-p)p}{MK+1}}} dp}$$

```
Print[
  "Normal approximation: " <> ToString[NIntegrate[p PDF[NormalDistribution[p, Sqrt[p * (1 - p) / (M K)]],
      Total[observations] / (M K)], {p, 0, 1}] / NIntegrate[
    PDF[NormalDistribution[p, Sqrt[p * (1 - p) / (M K)]], Total[observations] / (M K)], {p, 0, 1}]]];
```

```
Normal approximation: 0.00009995
```

It should be noted that $\lim_{MK \to \infty} \int_0^1 \frac{e^{-\frac{p M K}{2(1-p)}}}{\sqrt{2\pi}\sqrt{\frac{(1-p)p}{MK+1}}} dp = 1$. Therefore the above approximation reduces to

$$\tilde{f}(p \mid 0) = \int_0^1 p \frac{e^{-\frac{p M K}{2(1-p)}}}{\sqrt{2\pi}\sqrt{\frac{(1-p)p}{MK+1}}} dp$$

```
Print["Simplified normal approximation: " <> ToString[NIntegrate[
    p PDF[NormalDistribution[p, Sqrt[p * (1 - p) / (M K)]], Total[observations] / (M K)], {p, 0, 1}]]];
```

```
Simplified normal approximation: 0.00009994
```

We shall now prove this result formally. Let T be the number of past observations of zeroes.

■ Proof

$$\int_0^1 \frac{e^{-\frac{pT}{2(1-p)}}}{\sqrt{2\pi}\sqrt{\frac{(1-p)p}{T}}}\,dp$$

$$\frac{e^{\frac{T}{2}}}{\sqrt{\frac{2\pi}{T}}}\int_0^1 \frac{e^{-\frac{T}{2p}}}{\sqrt{(1-p)p}}\,dp$$

$$\frac{e^{\frac{T}{2}}}{\sqrt{\frac{2\pi}{T}}}\,\pi\,\frac{2}{\sqrt{\pi}}\int_{\frac{\sqrt{T}}{\sqrt{2}}}^{\infty} e^{-t^2}\,dt$$

$$\frac{e^{\frac{T}{2}}}{\sqrt{\frac{2\pi}{T}}}\,2\pi\,\frac{1}{\sqrt{\pi}}\int_{\frac{\sqrt{T}}{\sqrt{2}}}^{\infty} e^{-t^2}\,dt$$

see bounds in http : // www.johndcook.com / normalbounds.pdf

$$\frac{1}{2\sqrt{\pi}}\,\frac{t}{t^2+1}\,e^{-t^2} \le \frac{1}{\sqrt{\pi}}\int_t^{\infty} e^{-i^2}\,di \le \frac{1}{2\sqrt{\pi}}\,\frac{1}{t}\,e^{-t^2}$$

$$\frac{1}{2\sqrt{\pi}}\,\frac{\frac{\sqrt{T}}{\sqrt{2}}}{\frac{T}{2}+1}\,e^{-\frac{T}{2}} \le \frac{1}{\sqrt{\pi}}\int_{\frac{\sqrt{T}}{\sqrt{2}}}^{\infty} e^{-t^2}\,dt \le \frac{1}{2\sqrt{\pi}}\,\frac{1}{\frac{\sqrt{T}}{\sqrt{2}}}\,e^{-\frac{T}{2}}$$

$$\frac{e^{-T/2}\sqrt{T}}{2\sqrt{2\pi}\,\left(1+\frac{T}{2}\right)} \le \frac{1}{\sqrt{\pi}}\int_{\frac{\sqrt{T}}{\sqrt{2}}}^{\infty} e^{-t^2}\,dt \le \frac{e^{-T/2}}{\sqrt{2\pi}\,\sqrt{T}}$$

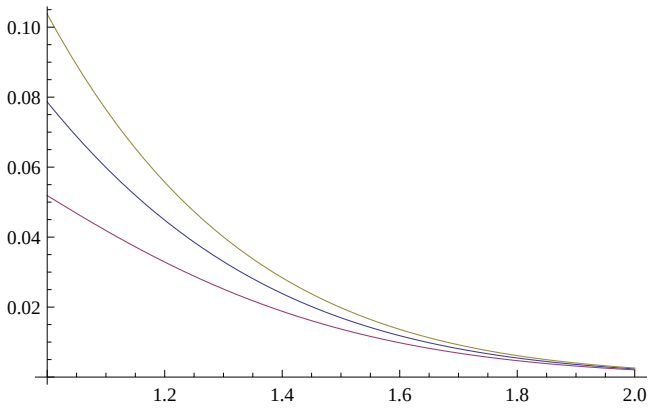
$$\frac{e^{\frac{T}{2}}}{\sqrt{\frac{2\pi}{T}}}\,2\pi\,\frac{e^{-T/2}}{\sqrt{2\pi}\,\sqrt{T}} = 1$$

$$\lim_{T\to\infty} \frac{e^{\frac{T}{2}}}{\sqrt{\frac{2\pi}{T}}}\,2\pi\,\frac{e^{-T/2}\sqrt{T}}{2\sqrt{2\pi}\,\left(1+\frac{T}{2}\right)} = 1$$

hence $\lim_{T\to\infty}\int_0^1 \frac{e^{-\frac{pT}{2(1-p)}}}{\sqrt{2\pi}\sqrt{\frac{(1-p)p}{T}}}\,dp=1\square$

We shall also provide graphs for the bounds in http : // www.johndcook.com/normalbounds.pdf.

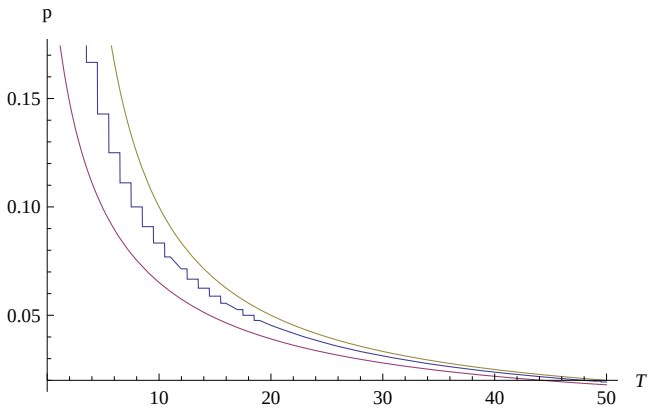
```
Plot[{ 1/Sqrt[π] Integrate[E^-i^2 di, {i, t, ∞}], 1/(2 Sqrt[π]) t/(t^2 + 1) E^-t^2, 1/(2 Sqrt[π]) 1/t E^-t^2}, {t, 1, 2}]
```



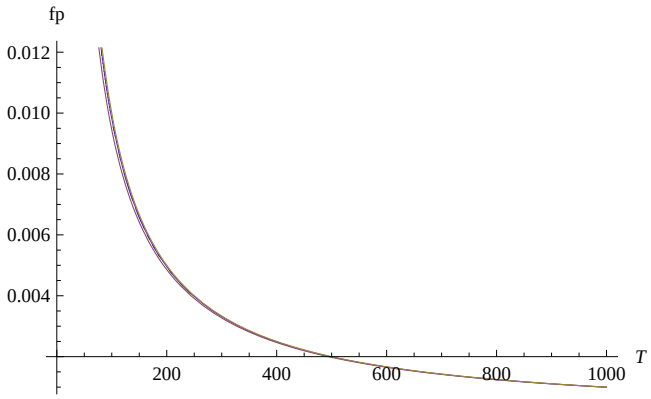
■ Hyperbolic decay

We shall now show that both $f(p \mid 0)$ and $\tilde{f}(p \mid 0)$ decay hyperbolically in the number of past zeroes observed.

```
Plot[
 {
  NIntegrate[i * PDF[BinomialDistribution[Round[T], i], 0], {i, 0, 1}] Round[T + 1],
  NIntegrate[i * PDF[NormalDistribution[i, Sqrt[i (1-i)/(T+1)]], 0], {i, 0, 1}]
  ---------------------------------------------------------------------------,
  NIntegrate[PDF[NormalDistribution[i, Sqrt[i*(1-i)/(T+1)]], 0], {i, 0, 1}]
  1 / T
 }, {T, 1, 50}, AxesLabel → {T, "p"}]
```



```
Plot[
 {
  NIntegrate[i * PDF[BinomialDistribution[Round[T], i], 0], {i, 0, 1}] Round[T + 1],
  NIntegrate[i * PDF[NormalDistribution[i, Sqrt[i (1-i)/(T+1)]], 0], {i, 0, 1}]
  ---------------------------------------------------------------------------,
    NIntegrate[PDF[NormalDistribution[i, Sqrt[i*(1-i)/(T+1)]], 0], {i, 0, 1}]
  1 / T
 }, {T, 1, 1000}, AxesLabel → {T, "fp"}]
```



Formally, this can be proved by considering that the posterior distribution $f(p \mid 0) = (T + 1) \int_0^1 p(1-p)^T dp$ can be integrated as follows.

$$(\text{Round}[T] + 1) \int_0^1 i * (1 - i)^{\text{Round}[T]} di = (\text{Round}[T] + 1) \frac{1}{2 + 3\,\text{Round}[T] + \text{Round}[T]^2} \simeq \frac{1}{\text{Round}[T]}$$