Fitting Theoretical Distributions

Jacob VanDrunen

December 13, 2019

Description

The purpose of this project is to help a scientist or statistician answer the question: given a set of data points, what is the most likely theoretical distribution that they could have come from? This problem has two sub-problems: (1) if it is distributed according to a given family of theoretical distributions, what is the most likely parametrization of the distribution, and (2) given the parametrized theoretical distribution, how closely does it "fit" the data? There are some methodological worries about, e.g., running certain statistical tests like the Kolmogorov-Smirnov test on a distribution that has already been parametrized based on the statistics of the data, but to get a rough estimate, this program would be acceptable.[1]

The three distributions that this program fits are the uniform, exponential, and normal distributions. The method of fitting distribution parameters is maximum likelihood estimation (MLE). For the uniform distribution, the MLE of the boundaries of the distribution is simply the minimum and maximum of the sampled data. For the exponential distribution, the MLE is either the mean of the data for the "scale" parameter β or equivalently the inverse of the mean for the "rate" parameter λ. For the normal distribution, the MLE for the mean is the sample mean, and the MLE for the variance is the variance taken without the loss of a degree of freedom.

The first test which the program runs to determine the fit of the model to the data is the Kolmogorov-Smirnov (K-S) test, which measures the greatest divergence of the cdf of the distribution and the edf of the sampled data. The sort of statistical test the program runs is one in which the null hypothesis is that the data comes from the given theoretical distribution. This means that the *higher* the p-value, the *better* the fit—as we are directly measuring $P(x|\theta)$. The actual likelihood that the data comes from the theoretical distribution cannot be determined without also specifying $P(\theta)$. This is left to the discretion of the user. A model with a significantly low p-value (which is defined arbitrarily by some α, traditionally 0.05) might be said to be sufficiently falsified by the K-S test.

The second test which the program runs is the Akaike information criterion (AIC). This is a test that is commonly used in Bayesian statistics to compare the relative likelihood of two models. It is defined (in its modified form to account for the sample size) as:

[1] Note also the screed printed at the beginning of the program's execution. Using p-values to form any sort of hard and fast conclusions about data is generally bad practice. Statistical analysis should always be accompanied by supplementary reasoning.

$$AIC = 2k - 2\log(\mathcal{L}) + \frac{2k^2 + 2k}{n - k - 1}$$

Where k is the number of parameters in the model, n is the sample size, and $\mathcal{L}$ is the likelihood of the model—defined here as the product of the likelihoods (measured with the pdf) of every observed datapoint on the given distribution (in the code, $\log(\mathcal{L})$ is calculated as the sum of the log-likelihoods using scipy's `logpdf` function). The resulting AIC can be used to compare the zoo of models tested on the data: the lower the AIC, the more likely the model is to minimize information loss, and to compute the relative likelihood of two models you take $P = e^{\frac{a-b}{2}}$ where a, b are the AICs of the two models. While this test does not produce p-values, I feel that it is a useful test for the stated goal of model comparison—even moreso than tests like K-S and χ^2.

As a final note, Dr. Hayes, thank you for teaching this class! It was by far the most fun class I took this quarter, and I learned a lot. I can tell that you have a certain nostalgic passion for this class, and you teach the material very well. I apologize that I only accomplished the bare minimum for this final project, but unhappily I have been very short on time this quarter due to other responsibilities. Have a good break!

Results

Generated Data

`data/uniform.txt`

```
KOLMOGOROV-SMIRNOV TEST
  p=0.61253  Uniform(75.03, 122.06)
  p=0.27377  Normal(mean=98.39, std=13.05)
  p=0.05000  ------------------------------------------------------------------
  p=0.00000  Exponential(rate=0.01)

AKAIKE INFORMATION CRITERION
        774  Uniform(75.03, 122.06)
        801  Normal(mean=98.39, std=13.05)
       1121  Exponential(rate=0.01)
```

Both K-S and AIC agree: the data is most likely uniformly distributed (this is correct), but on K-S only the exponential distribution is falsified by a significantly small p-value.

`data/exponential.txt`

```
KOLMOGOROV-SMIRNOV TEST
  p=0.87778  Exponential(rate=0.01)
  p=0.05000  ------------------------------------------------------------------
  p=0.00363  Normal(mean=101.72, std=107.73)
  p=0.00000  Uniform(0.04, 686.59)
```

```
AKAIKE INFORMATION CRITERION
       1128  Exponential(rate=0.01)
       1223  Normal(mean=101.72, std=107.73)
       1310  Uniform(0.04, 686.59)
```

Both K-S and AIC agree: the data is most likely exponentially distributed (this is also correct). On K-S both normal and uniform distributions are falsified with significantly small p-values.

```
data/normal.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.79449  Normal(mean=97.93, std=9.79)
  p=0.05000  ----------------------------------------------------------------
  p=0.00839  Uniform(75.27, 121.60)
  p=0.00000  Exponential(rate=0.01)

AKAIKE INFORMATION CRITERION
        744  Normal(mean=97.93, std=9.79)
        771  Uniform(75.27, 121.60)
       1120  Exponential(rate=0.01)
```

Both K-S and AIC agree: the data is most likely normally distributed (this is also correct). On K-S both normal and uniform distributions are falsified with significantly small p-values.

Provided Data

```
data/d1.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.36157  Exponential(rate=0.37)
  p=0.05000  ----------------------------------------------------------------
  p=0.00000  Normal(mean=2.73, std=2.70)
  p=0.00000  Uniform(0.00, 24.64)

AKAIKE INFORMATION CRITERION
      40064  Exponential(rate=0.37)
      48278  Normal(mean=2.73, std=2.70)
      64092  Uniform(0.00, 24.64)
```

For the first dataset, both K-S and AIC agree that it is most likely exponentially distributed. On K-S both other distributions are falsified with significantly small p-values.

```
data/d2.txt

KOLMOGOROV-SMIRNOV TEST
```

```
  p=0.27875  Uniform(2.72, 3.14)
  p=0.05000  ----------------------------------------------------------------
  p=0.00000  Normal(mean=2.93, std=0.12)
  p=0.00000  Exponential(rate=0.34)

AKAIKE INFORMATION CRITERION
     -17192  Uniform(2.72, 3.14)
     -13681  Normal(mean=2.93, std=0.12)
      41497  Exponential(rate=0.34)
```

For the second dataset, both K-S and AIC agree that it is most likely uniformly distributed. On K-S both other distributions are falsified with significantly small p-values.

```
data/d3.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.83406  Normal(mean=2.70, std=3.15)
  p=0.05000  ----------------------------------------------------------------
  p=0.00000  (Exponential undefined for domain: [-8.36, 14.75])
  p=0.00000  Uniform(-8.36, 14.75)

AKAIKE INFORMATION CRITERION
      51333  Normal(mean=2.70, std=3.15)
      62811  Uniform(-8.36, 14.75)
```

For the third dataset, both K-S and AIC agree that it is most likely normally distributed. The exponential distribution is certainly excluded because of the presence of negative numbers in the dataset. On K-S the uniform distribution is falsified with significantly small p-values.

```
data/d4.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.64413  Exponential(rate=0.36)
  p=0.05000  ----------------------------------------------------------------
  p=0.00000  Normal(mean=2.76, std=2.75)
  p=0.00000  Uniform(0.00, 19.29)

AKAIKE INFORMATION CRITERION
       4035  Exponential(rate=0.36)
       4865  Normal(mean=2.76, std=2.75)
       5923  Uniform(0.00, 19.29)
```

For the fourth dataset, both K-S and AIC agree that it is most likely exponentially distributed. On K-S both other distributions are falsified with significantly small p-values.

```
data/d5.txt
```

```
KOLMOGOROV-SMIRNOV TEST
  p=0.54180  Uniform(2.72, 3.14)
  p=0.05000  ----------------------------------------------------------------
  p=0.00131  Normal(mean=2.93, std=0.12)
  p=0.00000  Exponential(rate=0.34)

AKAIKE INFORMATION CRITERION
      -1717  Uniform(2.72, 3.14)
      -1395  Normal(mean=2.93, std=0.12)
       4152  Exponential(rate=0.34)
```

For the fifth dataset, both K-S and AIC agree that it is most likely uniformly distributed. On K-S both other distributions are falsified with significantly small p-values.

```
data/d6.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.85107  Normal(mean=2.67, std=3.25)
  p=0.05000  ----------------------------------------------------------------
  p=0.00000  (Exponential undefined for domain: [-6.79, 12.69])
  p=0.00000  Uniform(-6.79, 12.69)

AKAIKE INFORMATION CRITERION
       5198  Normal(mean=2.67, std=3.25)
       5943  Uniform(-6.79, 12.69)
```

For the sixth dataset, both K-S and AIC agree that it is most likely normally distributed. The exponential distribution is certainly excluded because of the presence of negative numbers in the dataset. On K-S the uniform distribution is falsified with significantly small p-values.

```
data/d7.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.43962  Exponential(rate=0.38)
  p=0.05000  ----------------------------------------------------------------
  p=0.00014  Normal(mean=2.62, std=2.75)
  p=0.00000  Uniform(0.02, 16.98)

AKAIKE INFORMATION CRITERION
        396  Exponential(rate=0.38)
        489  Normal(mean=2.62, std=2.75)
        570  Uniform(0.02, 16.98)
```

For the seventh dataset, both K-S and AIC agree that it is most likely exponentially distributed. On K-S the other two distributions are falsified with significantly small p-values.

```
data/d8.txt
```

```
KOLMOGOROV-SMIRNOV TEST
  p=0.41363  Normal(mean=2.93, std=0.11)
  p=0.24595  Uniform(2.72, 3.14)
  p=0.05000  ------------------------------------------------------------------
  p=0.00000  Exponential(rate=0.34)

AKAIKE INFORMATION CRITERION
       -169  Uniform(2.72, 3.14)
       -147  Normal(mean=2.93, std=0.11)
        419  Exponential(rate=0.34)
```

For the eighth dataset, the K-S test falsifies the exponential distribution, and declares the normal distribution more likely than the uniform distribution. The AIC disagrees, however, placing the uniform distribution as more likely. My hypothesis is that because of the small spread of the data, the uniform distribution's pdf is much higher, and thus the log-likelihood of the uniform distribution will be high for any point on the sampled distribution.

```
data/d9.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.95329  Normal(mean=2.49, std=3.21)
  p=0.05000  ------------------------------------------------------------------
  p=0.00210  Uniform(-6.79, 10.66)
  p=0.00000  (Exponential undefined for domain: [-6.79, 10.66])

AKAIKE INFORMATION CRITERION
        520  Normal(mean=2.49, std=3.21)
        576  Uniform(-6.79, 10.66)
```

For the ninth dataset, both K-S and AIC agree that it is most likely normally distributed. The exponential distribution is certainly excluded because of the presence of negative numbers in the dataset. On K-S the uniform distribution is falsified with significantly small p-values.

```
data/d10.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.93448  Exponential(rate=0.39)
  p=0.25679  Normal(mean=2.56, std=2.19)
  p=0.05000  ------------------------------------------------------------------
  p=0.02145  Uniform(0.19, 7.47)

AKAIKE INFORMATION CRITERION
         44  Exponential(rate=0.39)
         45  Uniform(0.19, 7.47)
         49  Normal(mean=2.56, std=2.19)
```

For the tenth dataset, both K-S and AIC agree that it is most likely exponentially distributed. However, while K-S falsifies uniform with significantly low p-value, AIC gives that the uniform is more likely than the normal distribution—in fact $e^{-\frac{1}{2}} \approx 0.61$ times as probable as the exponential distribution to minimize loss of information.

```
data/d11.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.86091  Uniform(2.75, 3.11)
  p=0.85297  Normal(mean=2.93, std=0.11)
  p=0.05000  ----------------------------------------------------------------
  p=0.00058  Exponential(rate=0.34)

AKAIKE INFORMATION CRITERION
        -14  Uniform(2.75, 3.11)
         -9  Normal(mean=2.93, std=0.11)
         47  Exponential(rate=0.34)
```

For the eleventh dataset, both K-S and AIC agree that it is most likely uniformly distributed. On K-S only the exponential distribution is falsified with significantly small p-values.

```
data/d12.txt

KOLMOGOROV-SMIRNOV TEST
  p=0.69295  Normal(mean=1.07, std=3.62)
  p=0.19878  Uniform(-6.79, 5.25)
  p=0.05000  ----------------------------------------------------------------
  p=0.00000  (Exponential undefined for domain: [-6.79, 5.25])

AKAIKE INFORMATION CRITERION
         55  Uniform(-6.79, 5.25)
         59  Normal(mean=1.07, std=3.62)
```

For the twelfth dataset, the exponential distribution is certainly falsified by the presence of negative numbers in the dataset. The K-S test delivers that the data is most likely normally distributed, but the AIC delivers that the data is most likely uniformly distributed, with the normally distributed model being $e^{-2} \approx 0.14$ times as probable to minimize information loss.

Code

```
#!/usr/bin/env python3


##  Fitting Theoretical Distributions
##  CS 115 Final Project
```

```
##
##  AUTHOR: Jacob VanDrunen [JVANDRUN]
##    DATE: December 12, 2019


import sys

import numpy as np
import scipy.stats as st


SCREED = '''
ARE P-VALUES BAD FOR YOUR HEALTH?

The ASA has issued the following 6 statements on the intended use of p-values.
Use the output of this program at your own risk.

 1. P-values can indicate how incompatible the data are with a specified
    statistical model.
 2. P-values do not measure the probability that the studied hypothesis is
    true, or the probability that the data were produced by random chance
    alone.
 3. Scientific conclusions and business or policy decisions should not be based
    only on whether a p-value passes a specific threshold.
 4. Proper inference requires full reporting and transparency.
 5. A p-value or statistical significance does not measure the size of an
    effect or the importance of a result.
 6. By itself, a p-value does not provide a good measure of evidence regarding
    a model or hypothesis.

For your Bayesian convenience, we have included the AIC measure for every
theoretically-fit distribution. As if that should make you more confident of
your decision.
'''


def mle_normal(values):
    # The MLE of the mean and variance of a normal distribution is the mean of
    # the observed distribution and the sample variance (that is, without the
    # loss of dof) of the observed distribution. scipy returns a "scale"
    # parameter which is equivalent to the standard deviation.
    params = st.norm.fit(values)
    return 'norm', params, 'Normal(mean={:.2f}, std={:.2f})'.format(*params)
```

```
def mle_exponential(values):
    # The MLE for the rate of an exponential distribution is the inverse of the
    # mean of the observed distribution.
    try:
        params = st.expon.fit(values, floc=0.)
        # scipy estimates a "scale" (beta) parameter, which is equal to
        #   1. / "rate" (lambda)
        rate = 1. / params[1]
        return 'expon', params, 'Exponential(rate={:.2f})'.format(rate)
    except Exception:
        return None, None, '(Exponential undefined for domain: [{:.2f}, {:.2f}])'.\
                           format(np.min(values), np.max(values))


def mle_uniform(values):
    # The MLE for the bounds of a uniform distribution are the max and min of
    # the observed distribution.
    params = st.uniform.fit(values)
    # scipy estimates "location" and "scale" parameters, where the domain is
    #   [location, location + scale]
    a = params[0]
    b = params[0] + params[1]
    return 'uniform', params, 'Uniform({:.2f}, {:.2f})'.format(a, b)


def ks_test(values, dist, params):
    # D-statistic determined by computing the supremum of the supremums of D+
    # and D-, the vertical distances (+ and -) between the edf of the sample
    # and the cdf of the distribution to test across all points
    D = st.kstest(values, dist, params, alternative='two-sided', mode='asymp').\
        statistic
    # p-value determined by computing the percentile of the D statistic
    # multiplied by the ratio given in class (sqrt(n) * D would also work in the
    # simpler case) on the Kolmogorov distribution
    p = 1. - st.kstwobign.cdf((np.sqrt(len(values)) +\
        (0.11 / np.sqrt(len(values))) + 0.12) * D)
    return p


def aic_test(values, dist, params):
    # Uses the modified Akaike information criterion to account for potentially
    # small sample sizes. The log-likelihood is calculated as the log of the
    # product of the probability density of every point in the observed data on
    # the proposed distribution (i.e. the sum of the logs of the same).
    k = len(params)
```

```
    ll = np.sum(getattr(st, dist).logpdf(values, *params))
    aic = 2.*k - 2.*ll + (2.*(k**2 + k))/(len(values) - k - 1.)
    return aic


def print_results(fits_ks, fits_aic):
    print('KOLMOGOROV-SMIRNOV TEST')
    for p_value, label in fits_ks:
        print('  p={:.05f}  {}'.format(p_value, label))
    print()
    print('AKAIKE INFORMATION CRITERION')
    for aic, label in fits_aic:
        print('  {:9d}  {}'.format(int(aic), label))


def distfit(values, dists):
    fits_ks = []
    fits_ks.append((0.05, '-'*67))
    fits_aic = []
    for estimator in dists:
        dist, params, label = estimator(values)
        if dist is not None:
            # NOTE: formally, this is very bad practice. We are performing our
            # statistical test against a distribution that has already been fit
            # with parameters derived from the empirical data. However, it seems
            # to me that we can get away with it here because our goal is to
            # find an ordinal comparison of distributions, thus the question we
            # are answering here is: given two potential families of
            # distributions, which one fits the data better? We are NOT asking:
            # given a distribution fit to the data, how well does it fit?
            # However, it means that our p-values should be treated with all the
            # more caution. Note also that the p-values are the reverse of how
            # they are typically treated: higher p-value means that there is a
            # higher chance that the null hypothesis (that the data is
            # distributed in the way proposed) is true.
            p_value = ks_test(values, dist, params)
            fits_ks.append((p_value, label))
            # AIC makes more sense, assuming that my comment above is correct.
            # AIC is a generic metric intended solely for model comparison (that
            # is, it does not have a simple interpretation outside of "this
            # model carries less risk than that one by some quantity." Lower is
            # better :)
            aic = aic_test(values, dist, params)
            fits_aic.append((aic, label))
        else:
```

```
            fits_ks.append((0., label))
    print_results(sorted(fits_ks, key=lambda t: t[0], reverse=True),
                  sorted(fits_aic, key=lambda t: t[0]))


if __name__ == '__main__':
    fit_dists = [
        mle_normal,
        mle_exponential,
        mle_uniform
    ]

    # Read from stdin if no files specified
    if len(sys.argv) == 1:
        lines = sys.stdin.readlines()
        print(SCREED)
        print()
        distfit([float(line.strip()) for line in lines], fit_dists)
        print()

    else:
        print(SCREED)
        for path in sys.argv[1:]:
            print()
            print(path)
            print()
            with open(path, 'r') as f:
                distfit([float(line.strip()) for line in f], fit_dists)
            print()
```