Statistical Inference

Fall 2016

Outline

Overview of Statistical Inference
Estimation and Prediction

Maximum Likelihood Theory

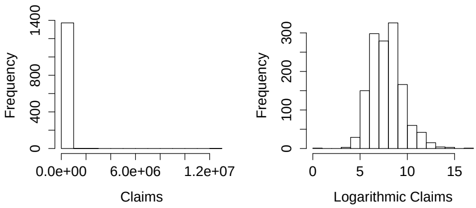
Overview of Statistical Inference

- A set of data (a **sample**) has been collected that is considered representative of a larger set (the **population**). This relationship is known as the **sampling frame**.
- Often, we can describe the distribution of the population in terms of a limited (finite) number of terms called **parameters**. These are referred to as *parametric distributions*. With **nonparametric** analysis, we do not limit ourselves to only a few parameters.
- The **statistical inference** goal is to say something about the (larger) population based on the observed sample (we "*infer*," not "*deduce*"). There are three types of statements:
    1. **Estimation**
    2. **Hypothesis Testing**
    3. **Prediction**

Wisconsin Property Fund

- Discuss ideas of statistical inference in the context of a sample from the Wisconsin Property Fund
- Specifically, consider 1,377 *individual* claims from 2010 experience (slightly different from the analysis of 403 average claims in Chapter 1)

	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Standard Deviation
Claims	1	788	2,250	26,620	6,171	12,920,000	368,030
Logarithmic Claims	0	6.670	7.719	7.804	8.728	16.370	1.683



Sampling Frame

- In statistics, a sampling frame **error** occurs when the sampling frame, the list from which the sample is drawn, is not an adequate approximation of the population of interest.
- For the property fund example, the sample consists of all 2010 claims
  - The population might be all claims that could have potentially occurred in 2010.
  - Or, it might be all claims that could potentially occur, such as in 2010, 2011, and so forth
- A sample must be a representative subset of a population, or “universe,” of interest. If the sample is not representative, taking a larger sample does not eliminate bias; you simply repeat the same mistake over again and again.

Sampling Frame II

- A sample should be a representative subset of a population, or "universe," of interest.
- Formally
  - We assume that the random variable X represents a draw from a population with distribution function F(.)
  - We make several such draws (n), each unrelated to one another (statistically independent)
  - Sometimes we say that $X_1, \ldots, X_n$ is a random sample (with replacement) from F(.)
  - Sometimes we say that $X_1, \ldots, X_n$ are identically and independently distributed (iid)

Describing the Population

- We think of the random variable X as a draw from the population with distribution function F(.)
- There are several ways to summarize F(.). We might consider the mean, standard deviation, 95th percentile, and so on.
    - Because these summary stats do not depend on a specific parametric reference, they are **nonparametric** summary measures.
- In contrast, we can think of logarithmic claims as normally distributed with mean μ and standard deviation σ, that is, claims have a *lognormal* distribution
- We will also look at the gamma distribution, with parameters α and θ, as a claims model
    - The normal, lognormal, and gamma are examples of **parametric** distributions.
    - The quantities μ, σ, α, and θ are known as *parameters*. When we know the parameters of a distribution family, then we have knowledge of the entire distribution.

Estimation

- Use θ to denote a summary of the population.
    - Parametric - It can be a parameter from a distribution such as μ or σ.
    - Nonparametric - It can also be a nonparametric summary such as the mean or standard deviation.
- Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be a function of the sample that provides proxy, or **estimate**, of θ. It is a function of the sample $X_1, \ldots, X_n$.
- In our property fund case,
    - 7.804 is a (nonparametric) estimate of the population expected logarithmic claim and 1.683 is an estimate of the corresponding standard deviation.
    - These are (parametric) estimates of the normal distribution for logarithmic claims
    - The estimate of the expected claim using the lognormal distribution is 10,106.8 (=$\exp(7.804 + 1.683^2/2)$).

Lognormal Distribution and Estimation

- Assume that claims follow a lognormal distribution, so that logarithmic claims follow the familiar normal distribution.
- Specifically, assume $\ln X$ has a normal distribution with mean μ and variance σ^2, sometimes denoted as $X \sim N(\mu, \sigma^2)$.
- For the property data, estimates are $\hat{\mu} = 7.804$ and $\hat{\sigma} = 1.683$. The "hat" notation is common. These are said to be **point estimates**, a single approximation of the corresponding parameter.
- Under general maximum likelihood theory (that we will do in a little bit), these estimates typically have a normal distribution for large samples.
  - Using notation, $\hat{\theta}$ has an approximate normal distribution with mean θ and variance, say, $\mathrm{Var}(\hat{\theta})$.
  - Take the square root of the variance and plug-in the estimate to define $se(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$. A **standard error** is an estimated standard deviation.
  - The next step in the mathematical statistics theory is to establish that $(\hat{\theta} - \theta)/se(\hat{\theta})$ has a t-distribution with "degrees of freedom" (a parameter of the distribution) equal to the sample size minus the dimension of θ.

Lognormal Distribution and Estimation II

- Assume that claims follow a lognormal distribution, so that logarithmic claims follow the familiar normal distribution.
- Under general maximum likelihood theory
  - $\hat{\theta}$ has an approximate normal distribution with mean θ and variance, say, $\mathrm{Var}(\hat{\theta})$.
  - Take the square root of the variance and plug-in the estimate to define $se(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$. A **standard error** is an estimated standard deviation.
  - $(\hat{\theta} - \theta)/se(\hat{\theta})$ has a t-distribution with "degrees of freedom" (a parameter of the distribution) equal to the sample size minus the dimension of θ.
  - As an application, we can invert this result to get a **confidence interval** for θ.
- A pair of statistics, $\hat{\theta}_1$ and $\hat{\theta}_2$, provide an interval of the form $[\hat{\theta}_1, \hat{\theta}_2]$ This interval is a $1-\alpha$ confidence interval for θ if $\Pr\left(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\right) \geq 1-\alpha$.
- For example, $\hat{\theta}_1 = \hat{\mu} - (t-value)\hat{\sigma}/\sqrt{n}$ and $\hat{\theta}_2 = \hat{\mu} + (t-value)\hat{\sigma}/\sqrt{n}$ provide a confidence interval for $\theta = \mu$. When $\alpha = 0.05$, $t-value \approx 1.96$.
- For the property fund, (7.715235, 7.893208) is a 95% confidence interval for μ.

Lognormal Distribution and Hypothesis Testing

An important statistical inference procedure involves verifying ideas about parameters.

- To illustrate, in the property fund, assume that mean logarithmic claims have historically been approximately been $\mu_0 = log(5000) = 8.517$. I might want to use 2010 data to see whether the mean of the distribution has changed. I also might want to test whether it has increased.
- The actual 2010 average was $\hat{\mu} = 7.804$. Is this a significant departure from $\mu_0 = 8.517$?
- One way to think about it is in terms of standard errors. The deviation is $(8.517 - 7.804)/(1.683/\sqrt{1377}) = 15.72$ standard errors. This is highly unlikely assuming an approximate normal distribution.

Lognormal Distribution and Hypothesis Testing II

- One hypothesis testing procedure begin with the calculation the test statistic $t - stat = (\hat{\theta} - \theta_0)/se(\hat{\theta})$. Here, θ_0 is an assumed value of the parameter.
- Then, one rejects the hypothesized value if the test statistic $t - stat$ is "unusual." To gauge "unusual," use the same t-distribution as introduced for confidence intervals.
- If you only want to know about a difference, this is known as a "two-sided" test; use the same $t - value$ as the case for confidence intervals.
- If you want to investigate whether there has been an increase (or decrease), then use a "one-sided" test.
- Another useful concept in hypothesis testing is the p-value, which is short hand for probability value. For a data set, a p-value is defined to be the smallest significance level for which the null hypothesis would be rejected.

Property Fund – Other Distributions

- For numerical stability and extensions to regression applications, statistical packages often work with transformed version of parameters
- The following estimates are from the **R** package **VGAM** (the `vglm` function)

Distribution	Parameter Estimate	Standard Error	t-stat
Gamma	10.190	0.050	203.831
	-1.236	0.030	-41.180
Lognormal	7.804	0.045	172.089
	0.520	0.019	27.303
Pareto	7.733	0.093	82.853
	-0.001	0.054	-0.016
GB2	2.831	1.000	2.832
	1.203	0.292	4.120
	6.329	0.390	16.220
	1.295	0.219	5.910

Likelihood Function

- Let $\mathrm{f}(\cdot;\boldsymbol{\theta})$ be the probability mass function if X is discrete or the probability density function if it is continuous.
- The likelihood is a function of the parameters ($\boldsymbol{\theta}$) with the data ($\mathbf{x}$) fixed rather than a function of the data with the parameters fixed.
- Define the *log-likelihood function*,

$$L(\boldsymbol{\theta}) = L(\mathbf{x};\boldsymbol{\theta}) = \ln \mathrm{f}(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \mathrm{f}(x_i;\boldsymbol{\theta}),$$

  evaluated at a realization $\mathbf{x}$.
- In the case of independence, the joint density function can be expressed as a product of the marginal density functions and, by taking logarithms, we can work with sums.

Example. Pareto Distribution

- Suppose that $X_1, \ldots, X_n$ represent a random sample from a single-parameter Pareto with cumulative distribution function:

$$\mathrm{F}(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

- In this case, the single parameter is $\theta = \alpha$.
- The corresponding probability density function is $\mathrm{f}(x) = 500^{\alpha}\alpha x^{-\alpha-1}$ and the logarithmic likelihood is

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln \mathrm{f}(x_i;\alpha) = n\alpha \ln 500 + n \ln \alpha - (\alpha+1)\sum_{i=1}^{n} \ln x_i.$$

Properties of Likelihood Functions

- One basic property of likelihood functions is:

$$\mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})\right) = \mathbf{0}$$

- The derivative of the log-likelihood function, $\partial L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, is called the *score function*.
- To see this,

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta})\right) &= \mathrm{E}\left(\frac{\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{f}(\mathbf{x}; \boldsymbol{\theta})}{\mathrm{f}(\mathbf{x}; \boldsymbol{\theta})}\right) = \int \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{f}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{y} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int \mathrm{f}(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{y} = \frac{\partial}{\partial \boldsymbol{\theta}} 1 = \mathbf{0}.
\end{aligned}$$

Properties of Likelihood Functions II

- Another basic property is:

$$\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}L(\boldsymbol{\theta})\right)+\mathrm{E}\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)=\mathbf{0}.$$

- With this, we can define the *information matrix*

$$\mathbf{I}(\boldsymbol{\theta})=\mathrm{E}\left(\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)=-\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}L(\boldsymbol{\theta})\right).$$

- In general

$$\frac{\partial}{\partial\boldsymbol{\theta}}L(\boldsymbol{\theta})=\frac{\partial}{\partial\boldsymbol{\theta}}\ln\prod_{i=1}^{n}\mathrm{f}(x_i;\boldsymbol{\theta})=\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ln\mathrm{f}(x_i;\boldsymbol{\theta}).$$

  has a large sample **normal distribution** with mean **0** and variance $\mathbf{I}(\boldsymbol{\theta})$.

Maximum Likelihood Estimators

- The value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_{MLE}$, that maximizes $\mathrm{f}(\mathbf{x}; \boldsymbol{\theta})$ is called the *maximum likelihood estimator*.
- Maximum likelihood estimators are values of the parameters $\boldsymbol{\theta}$ that are “most likely” to have been produced by the data.
- Because $\ln(\cdot)$ is a one-to-one function, we can also determine $\boldsymbol{\theta}_{MLE}$ by maximizing the log-likelihood function, $L(\boldsymbol{\theta})$.

Example. Course C/Exam 4. May 2000, 21. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with cumulative distribution function:

$$\mathrm{F}(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

Calculate the maximum likelihood estimate of the parameter α.

Instructor Notes

Example. Course C/Exam 4. May 2000, 21. You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with cumulative distribution function:

$$\mathrm{F}(x) = 1 - \left(\frac{500}{x}\right)^{\alpha}, \quad x > 500.$$

Calculate the maximum likelihood estimate of the parameter α.

Solution. With $n = 5$, the logarithmic likelihood is

$$L(\alpha) = \sum_{i=1}^{5} \ln \mathrm{f}(x_i;\alpha) = 5\alpha \ln 500 + 5 \ln \alpha - (\alpha+1)\sum_{i=1}^{5} \ln x_i.$$

Solving for the root of the score function yields

$$\frac{\partial}{\partial \alpha} L(\alpha) = 5 \ln 500 + 5/\alpha - \sum_{i=1}^{5} \ln x_i =_{set} 0 \Rightarrow \alpha_{MLE} = \frac{5}{\sum_{i=1}^{5} \ln x_i - 5 \ln 500} = 2.453.$$

Asymptotic Normality of Maximum Likelihood Estimators

- Under broad conditions, $\boldsymbol{\theta}_{MLE}$ has a large sample normal distribution with mean $\boldsymbol{\theta}$ and variance $(\mathbf{I}(\boldsymbol{\theta}))^{-1}$.
- $2\left(L(\boldsymbol{\theta}_{MLE}) - L(\boldsymbol{\theta})\right)$ has a chi-square distribution with degrees of freedom equal to the dimension of $\boldsymbol{\theta}$.
- These are critical results upon which much of estimation and hypothesis testing is based.

Example. Course C/Exam 4. Nov 2000, 13. A sample of ten observations comes from a parametric family $f(x,;\theta_1,\theta_2)$ with log-likelihood function

$$L(\theta_1,\theta_2) = \sum_{i=1}^{10} f(x_i;\theta_1,\theta_2) = -2.5\theta_1^2 - 3\theta_1\theta_2 - \theta_2^2 + 5\theta_1 + 2\theta_2 + k,$$

where k is a constant. Determine the estimated covariance matrix of the maximum likelihood estimator, $\hat{\theta}_1, \hat{\theta}_2$.

Instructor Notes

Example. Course C/Exam 4. Nov 2000, 13. A sample of ten observations comes from a parametric family $f(x,;\theta_1,\theta_2)$ with log-likelihood function

$$L(\theta_1,\theta_2)=\sum_{i=1}^{10} f(x_i;\theta_1,\theta_2)=-2.5\theta_1^2-3\theta_1\theta_2-\theta_2^2+5\theta_1+2\theta_2+k,$$

where k is a constant. Determine the estimated covariance matrix of the maximum likelihood estimator, $\hat{\theta_1},\hat{\theta_2}$.
Solution. The matrix of second derivatives is

$$\left(\begin{array}{cc} \frac{\partial^2}{\partial\theta_1^2}L & \frac{\partial^2}{\partial\theta_1\partial\theta_2}L \\ \frac{\partial^2}{\partial\theta_1\partial\theta_2}L & \frac{\partial^2}{\partial\theta_1^2}L \end{array}\right)=\left(\begin{array}{cc} -5 & -3 \\ -3 & -2 \end{array}\right)$$

Thus, the information matrix is:

$$\mathbf{I}(\theta_1,\theta_2)=-\mathrm{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}L(\boldsymbol{\theta})\right)=\left(\begin{array}{cc} 5 & 3 \\ 3 & 2 \end{array}\right)$$

and

$$\mathbf{I}^{-1}(\theta_1,\theta_2)=\frac{1}{5(2)-3(3)}\left(\begin{array}{cc} 2 & -3 \\ -3 & 5 \end{array}\right)=\left(\begin{array}{cc} 2 & -3 \\ -3 & 5 \end{array}\right).$$

Maximum Likelihood Estimation (MLE)

- Why use maximum likelihood estimation?
  - General purpose tool - works in many situations (data can be censored, truncated, include covariates, time-dependent, and so forth)
  - It is “optimal,” the best, in the sense that it has the smallest variance among the class of all unbiased estimators. (Caveat: for large sample sizes).
- A drawback: Generally, maximum likelihood estimators are computed iteratively, no closed-form solution.
  - For example, you may recall a “Newton-Raphson” iterative algorithm from calculus
  - Iterative algorithms require starting values. For some problems, the choice of a close starting value is critical.

MLE and Statistical Significance

One important type inference is to say whether a parameter estimate is "statistically significant"

- We learned earlier that $\boldsymbol{\theta}_{MLE}$ has a large sample normal distribution with mean $\boldsymbol{\theta}$ and variance $(\mathbf{I}(\boldsymbol{\theta}))^{-1}$.
- Look to the jth element of $\boldsymbol{\theta}_{MLE}$, say $\theta_{MLE,j}$.
- Define $se(\theta_{MLE,j})$, the standard error (estimated standard deviation) to be square root of the j diagonal element of $(\mathbf{I}(\boldsymbol{\theta})_{MLE})^{-1}$.
- To assess the hypothesis that θ_j is 0, we look at the rescaled estimate $t(\theta_{MLE,j}) = \theta_{MLE,j}/se(\theta_{MLE,j})$. It is said to be a t-statistic or t-ratio.
- Under this hypothesis, it has a t-distribution with degrees of freedom equal to the sample size minus the dimension of $\boldsymbol{\theta}_{MLE}$.
- For most actuarial applications, the t-distribution is very close to the (standard) normal distribution. Thus, sometimes this ratio is also known a z-statistic or "z-score."

MLE and Statistical Significance II

Assessing Statistical Significance

- If the t-statistic $t(\theta_{MLE,j})$ exceeds a cut-off (in absolute value), then the jth variable is said to be "statistically significant."
  - For example, if we use a 5% significance level, then the cut-off is 1.96 using a normal distribution approximation.
  - More generally, using a $100\alpha\%$ significance level, then the cut-off is a $100(1-\alpha/2)\%$ quantile from a t-distribution using degrees of freedom equal to the sample size minus the dimension of $\boldsymbol{\theta}_{MLE}$.
- Another useful concept in hypothesis testing is the p-value, shorthand for probability value.
  - For a data set, a p-value is defined as the smallest significance level for which the null hypothesis would be rejected.
  - The p-value is a useful summary statistic for the data analyst to report because it allows the reader to understand the strength of the deviation from the null hypothesis.

MLE and Model Validation

Another important type inference is to select a model from two choices, where one choice is a subset of the other

- Suppose that we have a (large) model and determine the maximum likelihood estimator, $\boldsymbol{\theta}_{MLE}$.
- Now assume that p elements in $\boldsymbol{\theta}$ are equal to zero and determine the maximum likelihood estimator over the remaining set. Call this estimator $\boldsymbol{\theta}_{Reduced}$
- The statistic, $LRT = 2\left(L(\boldsymbol{\theta}_{MLE}) - L(\boldsymbol{\theta}_{Reduced})\right)$, is called the likelihood ratio (a difference of the logs is the log of the ratio. Hence, the term “ratio.”)
- Under the hypothesis that the reduce model is correct, the likelihood ratio has a chi-square distribution with degrees of freedom equal to p, the number of variables set equal to zero.
- This allows us to judge which of the two models is correct. If the statistic LRT is large relative to the chi-square distribution, then we reject the simpler, reduced, model in favor of the larger one.

Information Criteria

- These statistics can be used when comparing several alternative models that are not necessarily nested. One picks the model that minimizes the criterion.
- *Akaike's Information Criterion*

$$AIC = -2 \times L(\boldsymbol{\theta}_{MLE}) + 2 \times (number\ of\ parameters)$$

  - The additional term $2\times$ $(number\ of\ parameters)$ is a penalty for the complexity of the model.
  - Other things equal, a more complex model means more parameters, resulting in a larger value of the criterion.
- *Bayesian Information Criterion*, defined as

$$BIC = -2 \times L(\boldsymbol{\theta}_{MLE}) + (number\ of\ parameters) \times \ln(number\ of\ observations)$$

  - This measure gives greater weight to the number of parameters.
  - Other things being equal, BIC will suggest a more parsimonious model than AIC.

Property Fund Information Criteria

- Both the AIC and BIC statistics suggest that the *GB2* is the best fitting model whereas gamma is the worst.

Distribution	AIC	BIC
Gamma	28,305.2	28,315.6
Lognormal	26,837.7	26,848.2
Pareto	26,813.3	26,823.7
GB2	26,768.1	26,789.0

Property Fund Fitted Distributions

- In this graph, black represents actual (smoothed) logarithmic claims
- Best approximated by green which is fitted GB2
- Pareto (purple) and Lognormal (lightblue) are also pretty good
- Worst are the exponential (in red) and gamma (in dark blue)