Introduction to the Theory of Statistics Part 2

PM522b

Meredith Franklin

Division of Biostatistics, University of Southern California

Slides 6, 2015

Outline

Topics Covered

1. Review of Convergence Concepts
   - Random sampling with large datasets
   - Convergence in probability
   - Almost sure convergence
   - Convergence in distribution
   - Central Limit Theorem
   - Slutsky's Theorem
2. Asymptotic Evaluations
   - Point Estimation: Consistency, Efficiency
   - Bootstrap
   - Robustness
   - Hypothesis Testing
   - Interval Estimation

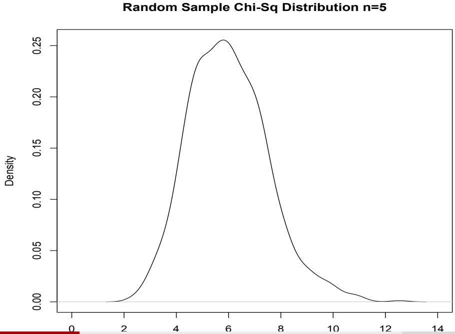
Random Sampling with Large Datasets

- We saw that estimates of population quantities from random samples rarely equal the true population quantity
- This is due to sampling variation (small samples result in unreliable representations of the population)
- We revisit the convergence behaviour of sample quantities as $n \to \infty$

Random Sampling with Large Datasets

Example: sampling and convergence

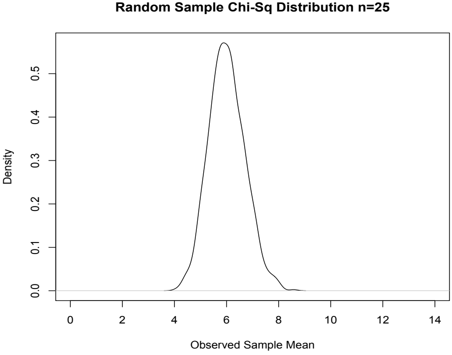
Take a random sample from χ^2_6 of size $n = 5$. Recall for a χ^2_k distribution, the mean $\mu = E(X_i) = k$, where k is the number of degrees of freedom.
By simulation, $x =$rchisq(5, df=6) gives $\bar{x}_5 = 4.87$. If we take another random sample, $\bar{x}_5 = 4.39$. If we do this 1,000 times, we can see the distribution of $\bar{x}_5$ for $X_1, X_2, ..., X_5 \sim \chi^2_6$



Random Sampling with Large Datasets

Example (con't): sampling and convergence

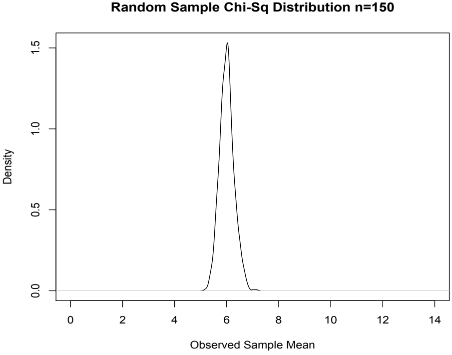
Take a random sample from χ^2_6 of size $n = 25$. By simulation, $x =$rchisq(25, df=6) gives $\bar{x}_{25} = 5.04$. If we take another random sample, $\bar{x}_{25} = 6.25$. If we do this 1,000 times, we can see the distribution of $\bar{x}_{25}$ for $X_1, X_2, ..., X_{25} \sim \chi^2_6$



Random Sampling with Large Datasets

Example (con't): sampling and convergence

Take a random sample from χ^2_6 of size $n = 150$. By simulation, $x =$rchisq(150, df=6) gives $\bar{x}_{150} = 5.91$. If we take another random sample, $\bar{x}_{150} = 6.13$. If we do this 1,000 times, we can see the distribution of $\bar{x}_{150}$ for $X_1, X_2, ..., X_{150} \sim \chi^2_6$



Random Sampling with Large Datasets

- We find that as $n \to \infty$ the sample mean, $\bar{X}_n$ narrows around the expected value (population mean)
- We know $E(X_i) = \mu$ by definition
- For the sample mean, $E(\bar{X}_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$
- We know $\mathrm{Var(X}_i) = \sigma^2$ by definition
- For the variance of the sample mean,

$$
\begin{aligned}
Var(\bar{X}_n) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Random Sampling with Large Datasets

Aside: Linear Transformation of Variance

For $a, b \in \Re$, $Var[a + bX] = b^2 Var[X]$ since:

$$\begin{aligned} Var[a + X] &= E[(a + X - E[a + X])^2] \\ &= E[(a + X - a - E[X])^2] \\ &= E[(X - E[X])^2] \\ &= Var[X] \end{aligned}$$

And

$$\begin{aligned} Var[bX] &= E[(bX - E[bX])^2] \\ &= E[(bX - bE[X])^2] \\ &= E[b^2(X - E[X])^2] \\ &= b^2 Var[X] \end{aligned}$$

Random Sampling with Large Datasets

- We see that the variance of the sample mean, $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$, has less variability than any of the individual random variables X_i being averaged, indicating that averaging decreases variation, so as $n \to \infty$, $\text{Var}(\bar{X}_n) \to 0$.
- If we repeat the experiment enough times we can make the variation around the sample mean infinitely small.

Convergence in Probability

This is the weaker of convergence types.

Definition of Convergence in Probability

For an iid sequence of random variables $X_1, X_2, ..., X_n$ and any positive constant ϵ

$$\lim_{n \to \infty} P(|\bar{X}_n - X| \geq \epsilon) = 0$$

or equivalently,

$$\lim_{n \to \infty} P(|\bar{X}_n - X| < \epsilon) = 1$$

Convergence in Probability

Convergence in probability is the type of convergence established by the weak law of large numbers (WLLN). The WLLN applies to the sample mean by the following:

Weak Law of Large Numbers

For an iid sequence of random variables $X_1, X_2, ..., X_n$ with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and sample mean $E(\bar{X}_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$

$$\bar{X}_n \xrightarrow{p} \mu \text{ when } n \to \infty$$

Convergence in probability of the mean of our sample, a random variable $\bar{X}_n$, to a constant μ requires only that μ exists.
The WLLN also states (via Markov's Inequality and Chebychev's Inequality):
For any positive constant ϵ, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$
Meaning that for any non-zero number, no matter how small, when the sample size (n) is large, there will be a very high probability that the average of the observations will be close to the expected value.

Convergence in Probability

Markov's Inequality

For a non-negative random variable X, $P(X \geq 0) = 1$ and positive constant ϵ

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

Proof: Consider $X \sim f(x_i) = P(X = x_i)$ is a discrete random variable (This also applies to cts r.v.)

$$\begin{aligned}
E(X) &= \sum_{i=0}^{\infty} x_i f(x_i) \\
&= \sum_{0}^{x_i < \epsilon} x_i f(x_i) + \sum_{x_i \geq \epsilon}^{\infty} x_i f(x_i) \\
&\geq \sum_{x_i \geq \epsilon}^{\infty} x_i f(x_i) \\
&\geq \epsilon \sum^{\infty} f(x_i) = \epsilon P(X \geq \epsilon)
\end{aligned}$$

Convergence in Probability

Chebychev's Inequality

This is a specific and useful result of Markov's Inequality
Substituting r.v. X with $\bar{X} - \mu$:

$$\begin{aligned} P(\bar{X} - \mu \geq \epsilon) &= P((\bar{X} - \mu)^2 \geq \epsilon^2) \\ &\leq \frac{E(\bar{X}_n - \mu)^2}{\epsilon^2} \\ &= \frac{Var(\bar{X}_n)}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \end{aligned}$$

As $n \to \infty$, $\frac{\sigma^2}{n\epsilon^2} \to 0$ resulting in:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

Convergence Almost Surely

We also distinguish convergence in probability and convergence almost surely

Convergence almost surely

Convergence in probability is defined as:

$$P(|X_n - X| \geq \epsilon) \to 0 \text{ when } n \to \infty$$

Convergence almost surely (stronger than convergence in probability) is defined as:

$$P(X_n \to X \text{ when } n \to \infty) = 1$$

Thus when X_n converges X with probability 1, X_n converges to X *almost surely*

$$X_n \overset{a.s.}{\to} X$$

Furthermore, by definition of the continuity of $h(\cdot)$, and for $\omega \in \Omega$ (the probability space Ω):

$$\text{as } n \to \infty,\ X_n(\omega) \overset{a.s.}{\to} X(\omega)$$
$$\text{as } n \to \infty,\ h(X_n(\omega)) \overset{a.s.}{\to} h(X(\omega))$$

Convergence in Probability

Strong Law of Large Numbers

For an iid sequence of random variables $X_1, X_2, ..., X_n$ with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and sample mean $E(\bar{X}_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$

$$\bar{X}_n \xrightarrow{p} \mu \text{ when } n \to \infty$$

The SLLN states:
For any positive constant ϵ, $P(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1$
Which in other words states that the sample mean almost surely converges to the expected value as $n \to \infty$
$P(\lim\limits_{n\to\infty} \bar{X}_n = \mu) = 1$
The SLLN can be interpreted as: with probability=1, the limit of $\bar{X}_n$ is μ

Convergence in Probability

- The law of averages is a common term often used to describe how "things tend to average out in the long run".
- Recall the experiment where a coin was tossed 10 times, and we observed 8 heads giving $\bar{X}_{10} = 0.8$, but if the coin was really fair we would have observed $\bar{X}_n = 0.5$.
- By the LLN we would remain confident that as n increased we would eventually see that $\bar{X}_n$ tended to 0.5.
- The conclusion of the law of averages is essentially the frequentist interpretation of probability.
- Through this we have mathematical justification for approximating statistics when they are unknown.

Convergence in Distribution

For an iid sequence of random variables $X_1, X_2, ..., X_n$
$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$
If F_{X_n} are the cdfs of X_n and F_X is the cdf of X then

$$X_n \xrightarrow{d} X \text{ when } n \to \infty$$

Convergence in Distribution

Some additional theorems:

- The sequence of random variables $X_1, X_2, ..., X_n$ that converges in probability to a random variable X also converges in distribution to X.
- The sequence of random variables $X_1, X_2, ..., X_n$ converges in probability to a constant μ if and only if the sequence also converges in distribution to μ.

This leads to the Central Limit Theorem:

Central Limit Theorem

For a sequence of random variables $X_1, X_2, ..., X_n$ having finite mean $\mu = E(X_i)$ and variance $\sigma^2 = Var(X_i) > 0$, we define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

$$P(Z_n \leq z) = F_n(z) \xrightarrow{d} F(z) \text{ as } n \to \infty$$

where $F(z)$ is the cdf of the standard normal distribution

Convergence in Distribution

- The CLT states that the behaviour of the average (or sum) of a large number of iid random variables will resemble the behaviour of a standard normal random variable
- This is true regardless of the distribution of the random variables being averaged
- How many random variables must be averaged? Depends on the distribution, but $n \geq 30$ is a general rule of thumb

Central Limit Theorem

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$
$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

Slutsky's Theorem

Slutsky's theorem is useful for defining joint distributions as long as one of the sequences of random variables converges to a constant

Slutsky's Theorem

For sequences of random variables $\{X_n\}$ and $\{Y_n\}$, if $X_n \xrightarrow{d} X$ in distribution and $Y_n \xrightarrow{p} a$ in probability (a is a constant), then:

$$Y_n + X_n \xrightarrow{d} X + a$$
$$Y_n X_n \xrightarrow{d} aX$$
$$X_n / Y_n \xrightarrow{d} X/a$$

A special case of Slutsky's theorem arises when two sequences of random variables converge to constants:
For sequences of random variables $\{X_n\}$ and $\{Y_n\}$, if $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$

$$Y_n + X_n \xrightarrow{p} a + b$$
$$Y_n X_n \xrightarrow{p} ab$$
$$X_n / Y_n \xrightarrow{p} a/b$$

Continuous Mapping Theorem

Continuous Mapping Theorem

For sequences of random variables $\{X_n\}$ where $X_n \xrightarrow{p} X$ in probability, and $h(\cdot)$ is a continuous function at X then

$$h(X_n) \xrightarrow{p} h(X)$$

Furthermore, if $X_n \xrightarrow{d} X$ in distribution then

$$h(X_n) \xrightarrow{d} h(X)$$

Continuous Mapping Theorem

Example Continuous Mapping Theorem

Using the above theorems, we can show that from an iid sample $X_1, ..., X_n$ with $E(X) = \mu$

$$\bar{X} \xrightarrow{p} \mu$$

and since $h(x) = x^2$ is a continuous function, it follows that

$$\bar{X}^2 \xrightarrow{p} \mu^2$$

We can also show that

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \xrightarrow{p} \sigma^2$$

So the sample standard deviation $S \xrightarrow{p} \sigma$ and

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \xrightarrow{d} N(0, 1)$$

Furthermore, $\frac{n(\bar{X}-\mu)^2}{S^2} \xrightarrow{d} \chi^2_1$

Asymptotic Evaluations

- So far in the context of estimation we have focused on procedures involving finite samples.
- Asymptotic theory is based on the assumption that we can keep collecting data, making our sample size infinite.
- Asymptotic properties describe the behavior of a procedure as the sample size becomes infinite; this is also called "large sample theory".
- The basic idea is that calculations simplify when sample sizes become infinite.
- Some techniques can only be applied under infinite sample size simplifications (e.g. bootstrap).

Asymptotic Evaluations

- In asymptotic theory, we concern ourselves with sequences of random variables and estimators.
- Many convergence concepts described above are familiar in the context of point estimation as $n \to \infty$
  - From intuition, consistency: as we collect more data in our sample, our estimator eventually gets close to the true parameter.
  - From intuition, efficiency: as we collect more data in our sample, our estimator eventually has minimum variance.
  - From intuition, asymptotic normality: as we collect more data in our sample, averages of random variables behave like normally distributed random variables.

Consistency

- Example: a coin is tossed n times, we have a binomial pdf for our random variable X with the probability of the toss resulting in heads being p
- The true parameter p is unknown, but the sample proportion X/n is an estimator of p
- As the number of tosses gets larger, X/n should get closer to the true value of p
- Following the properties of convergence in probability, in our example, we expect that as $n \to \infty$, $X/n \to p$
- Thus $\lim_{n\to\infty} P(|(X/n) - p| \leq \epsilon) \to 1$

Consistency

An estimator $\hat{\theta}$ is a consistent estimator of θ if for any positive number ϵ

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| \leq \epsilon) = 1$$

Consistency

As $n \to \infty$ the sample information becomes better and better and the estimator will be close to the target parameter with high probability. The general principle is as $n \to \infty$ an estimator converges to the "correct" value. If we observe $X_1, ..., X_n$ with pdf $f(X|\theta)$ then we can construct a sequence of estimators $W_n = W_n(X_1, ..., X_n)$, such as $\bar{X}_1 = X_1, \bar{X}_2 = (X_1 + X_2)/n, \bar{X}_3 = (X_1 + X_2 + X_3)/n$. This leads us to the formal definition of consistency:

Formal Definition of Consistency

A sequence of estimators W_n is consistent for the parameter θ if for every $\epsilon > 0$ and $\theta \in \Theta$:

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| < \epsilon) = 1$$

Equivalently,

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| \geq \epsilon) = 0$$

That is, a consistent sequence of estimators converges in probability to the parameter θ.

Consistency

- Recall Chebychev's Inequality:

$$P_\theta(|W_n - \theta| \geq \epsilon) \leq \frac{E_\theta[(W_n-\theta)^2]}{\epsilon^2}$$

- This allows us to state that a sequence of estimators W_n is consistent by:

$$\lim_{n\to\infty} E_\theta[(W_n - \theta)^2] = 0$$

- And from the definition of expectation, bias, and variance

$$E_\theta[(W_n - \theta)^2] = \text{Var}_\theta(W_n) + [B_\theta(W_n)]^2$$

- We can state that if W_n is a sequence of estimators of a parameter θ satisfying
  - $\lim_{n\to\infty} \text{Var}_\theta(W_n) = 0$
  - $\lim_{n\to\infty} \text{B}_\theta(W_n) = 0$
- Then W_n is a consistent sequence of estimators of θ

Consistency of MLEs

- MLEs are consistent estimators of their parameters, but to prove this we need to show that the underlying density/likelihood function satisfies certain regularity conditions

Regularity Conditions for consistency of MLEs

1. $X_1, ..., X_n$ are observed where $X_i \sim f(x|\theta)$ are iid
2. The parameter θ is identifiable; if $\theta \neq \theta^{'}$, then $f(x|\theta) \neq f(x|\theta^{'})$
3. The densities $f(x|\theta)$ have common support, and $f(x|\theta)$ is differentiable in θ
4. The parameter space Θ contains an open set of which the true parameter value θ_0 is an interior point. Sometimes stated as Θ being *compact*, and $\theta_0 \in Int(\Theta)$

Note: although these are stated in terms of the pdf, they equivalently apply to the likelihood

Consistency of MLEs

Under the regularity conditions, for $X_1, ..., X_n$ iid $f(x|\theta)$ with $L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$ and where $\hat{\theta}$ is the MLE of θ, and $\tau(\theta)$ is a continuous function of θ, for every $\epsilon > 0$ and $\theta \in \Theta$:

$$\lim_{n \to \infty} P_\theta(|\tau(\hat{\theta}) - \tau(\theta)| \geq \epsilon) = 0$$

That is, $\tau(\hat{\theta})$ is a consistent estimator of $\tau(\theta)$.

Asymptotic Efficiency

- Consistency is a relatively weak property and is considered necessary of all reasonable estimators
- Asymptotic efficiency deals with the asymptotic variance of estimators and helps us distinguish an estimator that is the "best"
- We need to calculate the asymptotic variance as follows: define the finite sample variance, then take the limit using a normalizing constant (so that the asymptotic variance doesn't go to 0)

Asymptotic Variance

For an estimator T_n, we calculate finite variance $Var(T_n)$ and then evaluate $\lim_{n\to\infty} k_n Var(T_n)$ where k_n is a normalizing constant used because in many instances $\lim_{n\to\infty} Var(T_n) \to 0$. The normalizing constant forces it to a non-zero limit.

Definition: Limiting Variance

If

$$\lim_{n\to\infty} k_n Var(T_n) = \tau^2 < \infty$$

where k_n is a sequence of constants then τ^2 is called the limiting variance. For example, for $\bar{X}_n$ iid $N(\mu, \sigma^2)$, if $T_n = \bar{X}_n$ then $\lim_{n\to\infty} nVar(T_n) = \sigma^2$ is the limiting variance of T_n.

There can be issues with the limiting variance if the limit approaches infinity (which it can do in cases such as $T_n = 1/\bar{X}_n$). In such cases, the approximate variance can be used (see CB section 5.5.4). Adopting this approach leads to the asymptotic variance.

Asymptotic Variance

Definition: Asymptotic Variance

For an estimator T_n suppose

$$k_n(T_n - \tau(\theta)) \xrightarrow{d} N(0, \sigma^2)$$

where k_n is a sequence of constants then σ^2 is called the asymptotic variance or variance of the limit distribution of T_n.

Efficiency

Efficiency relates to variance, and we show there is an optimal asymptotic variance related to the Cramer-Rao Lower Bound:

Efficient Estimators

A sequence of estimators W_n is asymptotically efficient for a parameter (function of a parameter) $\tau(\theta)$ if

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} N[0, \nu(\theta)]$$

and

$$\nu(\theta) = \frac{[\tau^{'}(\theta)]^2}{E_\theta[\frac{\partial}{\partial\theta}\log f(X|\theta)^2]}$$

The asymptotic variance of W_n attains the CRLB.

Efficiency of MLEs

As were necessary for showing the consistency of MLEs, regularity conditions are required for showing efficiency of MLEs. The two necessary regularity conditions are:

5. For every $X \in \mathcal{X}$ the density of $f(X|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ and $\int f(X|\theta)dx$ can be differentiated three times.
6. For any θ_0 (interior point) $\in \Theta$ there exists a positive number c and function $M(X)$ (both may depend on θ_0) such that

$$|\frac{\partial^3}{\partial\theta^3} log(f(x|\theta)| \leq M(X)$$
$$\forall X \in \mathcal{X}, \theta_0 - c < \theta < \theta_0 + c, \text{with} E_{\theta_0}[M(X)] < \infty$$

Efficiency of MLEs

With these additional regularity conditions (they apply to $f(X|\theta)$ and $L(\theta|X)$)

Asymptotic efficiency of MLEs

For$X_1, ..., X_n$ iid with $f(X|\theta)$, let $\hat{\theta}$ be the MLE for the parameter θ and $\tau(\theta)$ be a continuous function of θ

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \xrightarrow{d} N[0, \nu(\theta)]$$

Where $\nu(\theta)$ is the Cramer-Rao Lower Bound. So $\tau(\hat{\theta})$ is an asymptotically efficient estimator for $\tau(\theta)$. Note it is also a consistent estimator.

Efficiency, Asymptotic Variance and Information

See in-class notes.

Relative Efficiency

- It is possible to have more than one estimate of our target parameter, θ
- In such cases, we can use relative efficiency to assess which of the unbiased estimators has (relatively) smaller variance
- That is, if $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators, $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$ if $\mathrm{Var}(\hat{\theta}_2) > \mathrm{Var}(\hat{\theta}_1)$
- The efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is:

$$\mathrm{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

Asymptotic Relative Efficiency

- In an asymptotic context, we can use the asymptotic variance as a means of comparing estimators and determining efficiency
- Recall efficiency as defined by the ratio between the CRLB and variance

$$\text{eff}(\hat{\theta}) = \frac{I(\theta)^{-1}}{Var(\hat{\theta})}$$

Asymptotic Relative Efficiency

If two estimators W_n and V_n satisfy

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} N[0, \sigma^2_W]$$
$$\sqrt{n}[V_n - \tau(\theta)] \xrightarrow{d} N[0, \sigma^2_V]$$

the asymptotic relative efficiency of V_n with respect to W_n is ARE(V_n, W_n)

$$ARE(V_n, W_n) = \frac{\sigma^2_W}{\sigma^2_V}$$

Asymptotic Normality

- Another way to restate consistency is by $W_n - \theta \xrightarrow{p} 0$
- Since W_n is sequence of estimators of our parameter, $W_n - \theta$ is the error of estimation
- Consistency states that this error goes to zero
- However, we can examine this further and define the sampling distribution of $W_n - \theta$:

$$\sqrt{n}(W_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

  for some constant σ^2
- An estimate defined as above is consistent and **asymptotically normal**
- The asymptotic variance is σ^2
- Under asymptotic normality, estimators converge to the unknown parameter at rate $1/\sqrt{n}$

Asymptotic Normality and Consistency

In terms of MLEs, we showed that they are efficient and consistent. This is a redundant statement as an efficient estimator is only defined when the estimator is asymptotically normal, and asymptotic normality implies consistency.

$$\sqrt{n}\frac{(W_n-\mu)}{\sigma} \xrightarrow{d} Z,\ Z \sim N(0,1)$$

Applying Slutsky's Theorem:

$$(W_n-\mu) = \frac{\sigma}{\sqrt{n}}(\sqrt{n}\frac{(W_n-\mu)}{\sigma}) \rightarrow lim_{n\rightarrow\infty}\frac{\sigma}{\sqrt{n}}Z = 0$$

So $W_n - \mu \rightarrow 0$ in distribution. And convergence in distribution to a point is equivalent to convergence in probability, so W_n is a consistent estimator of μ.

Robustness

- We have assumed that the model we are working with is the correct one.
- From our 'correct' working model, we've derived estimators that are optimal.
- For example, from the likelihood approach we have seen that we get the best possible inference by achieving the CRLB.
- However, likelihood requires full specification of the probability structure. The MLE is efficient only if the specified model is correct.
- Robustness helps us answer the question: we've selected a model, but how do we know if the model we've selected is correct?

Robustness

From our model we want:

1. Optimal or near optimal efficiency.
2. Small deviations from model assumptions should only slightly impair the performance of the model.
3. Larger deviations from the model should not yield crazy results.

We can examine these three items with specific examples (e.g. Normal and Cauchy pdfs). Also, in terms of the 3rd item, we can define a breakdown value: the value where deviations from the model can cause catastrophic results.

Robustness

Is the sample mean robust?

1. $X_1, ..., X_n \sim N(\mu, \sigma^2)$, $\bar{X}$ has variance σ^2/n which attains the CRLB.
2. Investigate how $\bar{X}$ behaves under small deviations, δ.

$$f(x) = \begin{cases} N(\mu, \sigma^2) \text{ with probability } 1-\delta \\ f(x) \text{ with probability } \delta \end{cases} \tag{1}$$

   where $f(x)$ is a different distribution such as $f(X|\theta, \tau^2)$. Then, we find the variance of $\bar{X}$:

$$Var(\bar{X}) = (1-\delta)\frac{\sigma^2}{n} + \delta\frac{\tau^2}{n} + \frac{\delta(1-\delta)(\theta-\mu)^2}{n}$$

   If $\theta \approx \mu$ and $\tau^2 \approx \sigma^2$ then this is near optimal so $\bar{X}$ will be near optimal ($Var(\bar{X}) \to \frac{\sigma^2}{n}$). However, if $f(x)$ is Cauchy, $Var(\bar{X}) = \infty$ so we no longer have optimality.
3. Larger deviations from the model should not yield crazy results. If there is an outlying observation, we have to see effect of increasing that observation (consider $X_{(n)} = x$ where $x \to \infty$). What is the breakdown value?

Robustness

Breakdown Value

If we order our sample $X_{(1)}, ..., X_{(n)}$ and let T_n be a statistic for this sample, we define the breakdown value as $b, 0 \leq b \leq 1$ if for every $\epsilon > 0$

$$\lim_{X_{(\{(1-b)n\})\to\infty}} T_n < \infty \text{ and } \lim_{X_{(\{(1-b+\epsilon)n\})\to\infty}} T_n = \infty$$

Where the subscript of the limit identifies the percentile of X (see CB 5.4.2).

The breakdown value of the mean is b=0, meaning that if any fraction of the sample approaches infinity so does the mean.

The breakdown value of the median is b=0.5 as the median remains unchanged to changes in sample values. So the median is more robust.

But which one is better? It is a matter of robustness (median) vs. optimality (mean) which can be answered by looking at the asymptotic relative efficiency.

Robustness

Comparing robustness and optimality with ARE (median vs. mean)
Suppose $X_1, ..., X_n$ are an iid sample from a distribution with pdf f(x) and CDF F_x. Let M_n be the sample median and μ be the population median where $P(X_i \leq \mu) = 1/2$. From CB 10.2.3 it is shown that the limiting distribution of the median is

$$\sqrt{n}(M_n - \mu) \rightarrow N(0, 1/[2f(\mu)]^2)$$

Using this asymptotic variance we can look at the ARE of $\bar{X}$ with M_n for the normal, logistic and double exponential distributions. Note all three distributions are symmetric so the population mean equals the population median.
Exercise 10.23 done in class.

Hypothesis Testing

The asymptotic distribution of the likelihood ratio test is very useful, particularly when the formula for the test statistic $\lambda(x)$ is complicated and it is difficult to find its sampling distribution. Recall:

$$\lambda(x) = \lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)} = \frac{\sup_{\theta\in\Theta_0} L(\theta|x_1, ..., x_n)}{\sup_{\theta\in\Theta} L(\theta|x_1, ..., x_n)}$$

has an explicit form for the critical region

$$R = \{x_1, ..., x_n : \lambda(x) \leq k\}$$

with k chosen so that α-level test is

$$\sup_{\theta\in\Theta_0} P[\lambda(x) \leq k|\theta \in \Theta_0] \leq \alpha$$

Hypothesis Testing

For the hypothesis test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ where $\hat{\theta}$ is the MLE (satisfying the regularity conditions), then under H_0 as $n \to \infty$

$$-2 \log \lambda(X) \xrightarrow{d} \chi_1^2$$

Hypothesis Testing

Please note there is a typo in CB (Theorem 10.3.1)

- For the asymptotic distribution of the LRT testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ on p.489

Proof: The Taylor's Series expansion for $l(\theta|x)$ around $\hat{\theta}$ gives

$$l(\theta|x) = l(\hat{\theta}|x) + l'(\hat{\theta}|x)(\theta - \hat{\theta}) + l''(\hat{\theta}|x)\frac{(\theta - \hat{\theta})^2}{2!} + \ldots$$

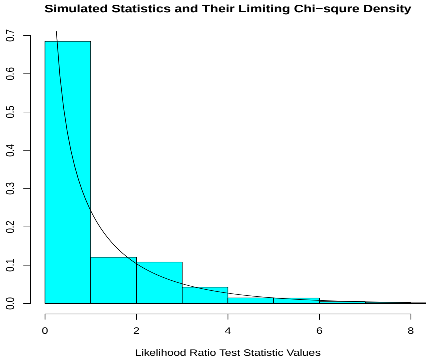
Now substitute the expansion for $l(\theta_0|x)$ in
$-2\log\lambda(x) = -2l(\theta_0|x) + 2l(\hat{\theta}|x)$ and get

$$-2\log\lambda(x) \approx -l''(\hat{\theta}|x)(\theta_0 - \hat{\theta})^2,$$

where we use the fact that $l'(\hat{\theta}|x) = 0$. Since $-l''(\hat{\theta}|x)$ is the observed information $\hat{I}_n(\hat{\theta})$ and $\frac{1}{n}\hat{I}_n(\hat{\theta}) \to I(\theta_0)$ it follows from Theorem 10.1.12 and Slutsky's Theorem (5.5.17) that $-2\log\lambda(\mathbf{X}) \to \chi^2_1$

Hypothesis Testing

Similar to CB Figure 10.3.1, we can visualize the asymptotic properties of the test statistic. Here we show the values of $-2\log\lambda(X)$ for the binomial distribution along with the pdf of χ^2_1



Hypothesis Testing

Another means of asymptotic hypothesis testing is based on the property of estimators having a normal distribution. For instance, W_n (e.g. the MLE) will have the following convergence

$$\frac{(W_n - \theta)}{\sigma_n} \xrightarrow{d} N(0,1)$$

This test is called the Wald test. Often the variance is unknown, so we use the asymptotic variance in the denominator of the LHS of this equation.

Confidence Intervals

Approximate Maximum Likelihood Intervals

- The confidence intervals we examined before are called "exact" as they require knowledge of the sampling distribution.
- An alternate method of constructing CI is based on large sample theory.
- This can be applied to maximum likelihood estimators.
- As we discussed previously by invariance, if $\hat{\theta}$ is the MLE of θ then $t(\hat{\theta})$ is the MLE of $t(\theta)$
- For large samples ($n \geq 35$) we can use the following as our pivot in determining confidence intervals for MLEs:

$$Z = \frac{t(\hat{\theta}) - t(\theta)}{\sqrt{[\frac{\partial t(\theta)}{\partial \theta}]^2 / nE[-\frac{\partial^2 logL(x|\theta)}{\partial \theta^2}]}}$$

- Recall the Cramer-Rao lower bound in its general form ($Var(t) \geq [\phi']^2/I(\theta)$) and the the denominator of Z where $[\phi']^2$ is $[\frac{\partial t(\theta)}{\partial \theta}]^2$
- $Z \sim N(0,1)$ by Slutsky's theorem and aymptotic properties of MLEs

Confidence Intervals

Example:Confidence Interval for MLE

Suppose we want to find a $100\%(1-\alpha)$ confidence interval for the variance of a Bernoulli random variable $(\theta(1-\theta))$.

By invariance the MLE of $t(\theta)=\theta(1-\theta)$ is $t(\hat{\theta})=\hat{\theta}(1-\hat{\theta})$. For $t(\theta)=\theta-\theta^2$ we have $\frac{\partial t(\theta)}{\partial\theta}=1-2\theta$

$$
\begin{aligned}
f(x|\theta) &= L(\theta|x) = \theta^x(1-\theta)^{1-x} \\
\log L(\theta|x) &= x\log\theta + (1-x)\log(1-\theta) \\
\frac{\partial \log L(\theta|x)}{\partial\theta} &= \frac{x}{\theta} + \frac{1-x}{1-\theta} \\
\frac{\partial^2 \log L(\theta|x)}{\partial\theta^2} &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \\
E[-\frac{\partial^2 \log L(\theta|x)}{\partial\theta^2}] &= E[\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}
\end{aligned}
$$

Confidence Intervals

Example:Confidence Interval for MLE con't

Putting everything together and using $Z = \frac{t(\hat{\theta})-t(\theta)}{\sqrt{[\frac{\partial t(\theta)}{\partial \theta}]^2/nE[-\frac{\partial^2 logL(x|\theta)}{\partial \theta^2}]}}$ as our pivotal quantity,

$$t(\hat{\theta}) \pm z_{\alpha/2}\sqrt{[\frac{\partial t(\theta)}{\partial \theta}]^2/nE[-\frac{\partial^2 logL(x|\theta)}{\partial \theta^2}]}$$

$$\hat{\theta}(1-\hat{\theta}) \pm z_{\alpha/2}\sqrt{(1-2\theta)^2/n[\frac{1}{\theta(1-\theta)}]}$$

$$\hat{\theta}(1-\hat{\theta}) \pm z_{\alpha/2}\sqrt{(1-2\hat{\theta})^2/n[\frac{1}{\hat{\theta}(1-\hat{\theta})}]}$$

Confidence Intervals

- An even simpler approximation for MLEs is often used:
$$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{E[I(\hat{\theta})]}}$$
- If n is large enough, the true coverage of this approximate interval will be very close to α
- $I(\theta) = E[-\frac{\partial^2 logL(x|\theta)}{\partial \theta^2}]$ is the expected information number.
$E[I(\hat{\theta})] = I(\hat{\theta}) = -\frac{\partial^2 logL(x|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}}$ is the observed information number.
- We use the observed information number and the approximation $Var(\hat{\theta}) \approx I(\hat{\theta})^{-1}$ to construct the approximate confidence interval on our MLE $\hat{\theta}$
- This becomes particularly useful when we need to use Newton's algorithm to estimate the Hessian, as the diagonal elements provide the information needed $I(\hat{\theta})$

Resampling Methods

- Resampling consists of a variety of methods that rely on repeated sampling rather than classical parameteric tests that compare observed statistics to theoretical sampling distributions.
- A computer is used to generate a large number of simulated samples.
- Samples are drawn (with replacement) from an existing sample of data, not from a theoretically defined distribution. The distribution is unknown, but the goal is to learn about the process (or distribution) that underlies the sample.
- Resampling methods include bootstrap, jacknife, randomization tests, and cross validation.
- Monte Carlo simulation is not the same as a resampling method because it involves generating a large number of samples from an assumed distribution (or model).

Bootstrap

- The boostrap method was introduced in 1979 by Efron
- The process of bootstrap is:
  - Start with an observed sample of size N
  - Generate a simulated sample of size N by drawing observations from the observed sample independently and with replacement
  - Calculate the statistic of interest
  - Repeat this many times (1,000+)
  - Have a distribution of the calculated statistic which is treated as an estimate of the population distribution of that statistic
- Resampling must be done with replacement, otherwise every simulated sample of size N would be identical to each other and the same as the original sample.
- Resampling with replacement means that some values may be sampled more frequently if they appear more often in the original sample.

Bootstrap Standard Errors

We generate information about estimators through resampling. In the case of the bootstrap, we re-sample with replacement (also called the non-parametric bootstrap). Recall that resampling with replacement results in

$$\binom{n+n-1}{n}$$

distinct samples, but they are not equiprobable. The n^n samples that are equally likely are trated as a random sample, though. For the ith resample, the mean is calculated as $\bar{x}_i^*$. The variance of this sample mean is:

$$Var^*(\bar{X}) = \frac{1}{n^n - 1} \sum_{i=1}^{n^n} (\bar{x}_i^* - \bar{\bar{x}}^*)^2$$

where

$$\bar{\bar{x}}^* = \frac{1}{n^n} \sum_{i=1}^{n^n} \bar{x}_i^*$$

is the mean of the re-samples.

Bootstrap Standard Errors

The advantage of the bootstrap and using this equation for the variance (the square root is the bootstrap standard error) is when there are large samples the delta method is applicable and we can use the asymptotic variance formula (with convergence in distribution to the normal). Specifically,

$$Var^*(\hat{\theta}) = \frac{1}{n^n - 1} \sum_{i=1}^{n^n} (\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2$$

where

$$\bar{\hat{\theta}}^* = \frac{1}{n^n} \sum_{i=1}^{n^n} \hat{\theta}_i^*$$

is the mean of the resamples.

Bootstrap Standard Errors

In the case of the binomial distribution, we the bootstrap binomal variance:

$$Var^*(\hat{p}(1-\hat{p})) = \frac{1}{n^n - 1}\sum_{i=1}^{n^n}(\hat{p}(1-\hat{p})^*_i - \overline{\hat{p}(1-\hat{p})^*})^2$$

Typically n^n is a very large number when we have a dataset with more than 15 observations. In this case, we don't enumerate all possible samples, but we select B re-samples (or bootstrap samples) and calculate

$$Var^*_B(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^*_i - \bar{\hat{\theta}}^*)^2$$

Parametric Bootstrap

The examples shown at the beginning of these set of slides used a 'plug-in' method, which is analogous to the parametric bootstrap.
Suppose we have a sample $X_1, ..., X_n$ with pdf $f(x|\theta)$ where θ may be a vector of parameters. We can estimate θ with $\hat{\theta}$ (e.g. the MLE) and draw samples from the distribution of $\hat{\theta}$.
In this case, the samples are not resamples of the data, but rather actual random samples from $f(x|\hat{\theta})$.
Parametric bootstrap can be considered a special case of the Monte Carlo method.

Bootstrap in R

The R package `boot` implements bootstrap methods. For example, to generate the bootstrap estimate of the sample mean we first define:

```
mean.boot <- function(x,index) {
mean(x[index])
}
```

Then we can call `boot` on this function as follows:

```
boot.mean<-boot(dat,mean.boot,1000)
```

And finally, take the standard deviation of the bootstrapped means

```
sd(boot.mean$t)
```

We can also use this to construct a confidence interval,

```
boot.ci(boot.mean, type = "norm")
```