Large sample tools[1]

STA442/2101 Fall 2018

[1] See last slide for copyright information.

Background Reading: Davison's *Statistical models*

- See Section 2.2 (Pages 28-37) on convergence.
- Section 3.3 (Pages 77-90) goes more deeply into simulation than we will. At least skim it.

Overview

Sample Space Ω, $\omega \in \Omega$

- Ω is a set, the underlying sample space.
- It could literally be the universe of websites from which we intend to sample.
- $\mathcal{F}$ is a class of subsets of Ω.
- If could be the class of all subsets (if Ω is countable).
- There is a probability measure $\mathcal{P}$ defined on the elements of $\mathcal{F}$.
- Maybe each website is equally likely to be chosen (with replacement).

Random variables are functions from Ω into the set of real numbers

$$Pr\{X \in B\} = Pr(\{\omega \in \Omega : X(\omega) \in B\})$$

Random Sample $X_1(\omega), \ldots, X_n(\omega)$

- $T = T(X_1, \ldots, X_n)$
- $T = T_n(\omega)$
- Let $n \to \infty$ to see what happens for large samples

Modes of Convergence

- Almost Sure Convergence
- Convergence in Probability
- Convergence in Distribution

Almost Sure Convergence

We say that T_n converges *almost surely* to T, and write $T_n \stackrel{a.s.}{\to} T$ if

$$Pr\{\omega : \lim_{n\to\infty} T_n(\omega) = T(\omega)\} = 1.$$

- Acts like an ordinary limit, except possibly on a set of probability zero.
- All the usual rules apply.
- Called convergence with probability one or sometimes strong convergence.
- In this course, convergence will usually be to a constant.

$$Pr\{\omega : \lim_{n\to\infty} T_n(\omega) = c\} = 1.$$

Strong Law of Large Numbers

Let $X_1, \ldots, X_n$ be independent with common expected value μ.

$$\overline{X}_n \stackrel{a.s.}{\to} E(X_i) = \mu$$

The only condition required for this to hold is the existence of the expected value.

Probability is long run relative frequency

- Statistical experiment: Probability of “success” is θ.
- Carry out the experiment many times independently.
- Code the results $X_i = 1$ if success, $X_i = 0$ for failure, $i = 1, 2, \ldots$

Sample proportion of successes converges to the probability of success

Recall $X_i = 0$ or 1.

$$\begin{aligned} E(X_i) &= \sum_{x=0}^{1} x \, Pr\{X_i = x\} \\ &= 0 \cdot (1-\theta) + 1 \cdot \theta \\ &= \theta \end{aligned}$$

Relative frequency is

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n \stackrel{a.s.}{\rightarrow} \theta$$

Simulation

Using pseudo-random number generation by computer

- Estimate almost any probability that's hard to figure out
- Statistical power
- Weather model
- Performance of statistical methods
- Need confidence intervals for estimated probabilities.

Estimating power by simulation

Example: Bernoulli random sampling.

Recall the two test statistics for testing $H_0 : \theta = \theta_0$:

- $Z_1 = \frac{\sqrt{n}(\overline{Y}-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$
- $Z_2 = \frac{\sqrt{n}(\overline{Y}-\theta_0)}{\sqrt{\overline{Y}(1-\overline{Y})}}$

When $\theta \neq \theta_0$, calculating $P\{|Z_2| > z_{\alpha/2}\}$ can be challenging.

Strategy for estimating power by simulation

- Generate a large number of random data sets under the alternative hypothesis.
- For each data set, test H_0.
- Estimated power is the proportion of times H_0 is rejected.
- How accurate is the estimate?
- $\widehat{p} \ \pm \ z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{m}}$

Testing $H_0 : \theta = 0.50$ when true $\theta = 0.60$ and $n = 100$

Power of Z_1 was about 0.52

$$Z_2 = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\overline{Y}(1 - \overline{Y})}}$$

```
> # Power by simulation
> set.seed(9999)
> m = 10000 # Monte Carlo sample size
> theta=0.60; theta0 = 1/2; n = 100
> Ybar = rbinom(m,size=n,prob=theta)/n  # A vector of length m
> Z2 = sqrt(n)*(Ybar-theta0)/sqrt(Ybar*(1-Ybar)) # Another vector of le
> power = length(Z2[abs(Z2>1.96)])/m; power

[1] 0.5394
```

Margin of error for estimated power

Confidence interval for an estimated probability was

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

```
# How about a 99 percent margin of error
> a = 0.005; z = qnorm(1-a)
> merror = z * sqrt(power*(1-power)/m); merror
[1] 0.0128391

> Lower = power - merror; Lower
[1] 0.5265609
> Upper = power + merror; Upper
[1] 0.5522391
```

Recall the Change of Variables formula: Let $Y = g(X)$

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

Or, for discrete random variables

$$E(Y) = \sum_y y\, p_Y(y) = \sum_x g(x)\, p_X(x)$$

This is actually a big theorem, not a definition.

Applying the change of variables formula

To approximate $E[g(X)]$

Simulate $X_1, \ldots, X_n$ from the distribution of X. Calculate

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n g(X_i) &= \frac{1}{n}\sum_{i=1}^n Y_i \stackrel{a.s.}{\to} E(Y) \\ &= E(g(X)) \end{aligned}$$

So for example

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \stackrel{a.s.}{\rightarrow} E(X^k)$$

$$\frac{1}{n}\sum_{i=1}^{n} U_i^2 V_i W_i^3 \stackrel{a.s.}{\rightarrow} E(U^2 V W^3)$$

That is, sample moments converge almost surely to population moments.

Approximate an integral: $\int_{-\infty}^{\infty} h(x)\,dx$

Where $h(x)$ is a nasty function.

Let $f(x)$ be a density with $f(x) > 0$ wherever $h(x) \neq 0$.

$$\begin{aligned}
\int_{-\infty}^{\infty} h(x)\,dx &= \int_{-\infty}^{\infty} \frac{h(x)}{f(x)} f(x)\,dx \\
&= E\left[\frac{h(X)}{f(X)}\right] \\
&= E[g(X)],
\end{aligned}$$

So

- Sample $X_1, \ldots, X_n$ from the distribution with density $f(x)$
- Calculate $Y_i = g(X_i) = \frac{h(X_i)}{f(X_i)}$ for $i = 1, \ldots, n$
- Calculate $\overline{Y}_n \stackrel{a.s.}{\rightarrow} E[Y] = E[g(X)]$
- Confidence interval for $\mu = E[Y]$ is routine.

Convergence in Probability

We say that T_n converges *in probability* to T, and write $T_n \stackrel{P}{\to} T$

if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\{\omega : |T_n(\omega) - T(\omega)| < \epsilon\} = 1$$

For us, convergence will usually be to a constant:

$$\lim_{n\to\infty} P\{|T_n - c| < \epsilon\} = 1$$

Convergence in probability (say to c) means no matter how small the interval around c, for large enough n (that is, for all $n > N_1$) the probability of getting that close to c is as close to one as you like.
We will seldom use the definition in this class.

Weak Law of Large Numbers

$$\overline{X}_n \xrightarrow{p} \mu$$

- Almost Sure Convergence implies Convergence in Probability
- Strong Law of Large Numbers implies Weak Law of Large Numbers

Consistency

$T = T(X_1, \ldots, X_n)$ is a statistic estimating a parameter θ

The statistic T_n is said to be *consistent* for θ if $T_n \stackrel{P}{\rightarrow} \theta$ for all θ in the parameter space.

$$\lim_{n \rightarrow \infty} P\{|T_n - \theta| < \epsilon\} = 1$$

The statistic T_n is said to be *strongly consistent* for θ if $T_n \stackrel{a.s.}{\rightarrow} \theta$.

Strong consistency implies ordinary consistency.

Consistency is great but it's not enough.

- It means that as the sample size becomes indefinitely large, you probably get as close as you like to the truth.
- It's the least we can ask. Estimators that are not consistent are completely unacceptable for most purposes.

$$T_n \stackrel{a.s.}{\rightarrow} \theta \Rightarrow U_n = T_n + \frac{100,000,000}{n} \stackrel{a.s.}{\rightarrow} \theta$$

Consistency of the Sample Variance

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

By SLLN, $\overline{X}_n \stackrel{a.s.}{\to} \mu$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \stackrel{a.s.}{\to} E(X^2) = \sigma^2 + \mu^2$.

Because the function $g(x, y) = x - y^2$ is continuous,

$$\widehat{\sigma}_n^2 = g\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2, \overline{X}_n\right) \stackrel{a.s.}{\to} g(\sigma^2 + \mu^2, \mu) = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

Convergence in Distribution

Sometimes called *Weak Convergence*, or *Convergence in Law*

Denote the cumulative distribution functions of $T_1, T_2, \ldots$ by $F_1(t), F_2(t), \ldots$ respectively, and denote the cumulative distribution function of T by $F(t)$.

We say that T_n converges *in distribution* to T, and write $T_n \stackrel{d}{\rightarrow} T$ if for every point t at which F is continuous,

$$\lim_{n\to\infty} F_n(t) = F(t)$$

Again, we will seldom use this definition directly.

Univariate Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from a distribution with expected value μ and variance σ^2. Then

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \stackrel{d}{\rightarrow} Z \sim N(0,1)$$

Connections among the Modes of Convergence

- $T_n \xrightarrow{a.s.} T \Rightarrow T_n \xrightarrow{p} T \Rightarrow T_n \xrightarrow{d} T.$
- If a is a constant, $T_n \xrightarrow{d} a \Rightarrow T_n \xrightarrow{p} a.$

Sometimes we say the distribution of the sample mean is approximately normal, or asymptotically normal.

- This is justified by the Central Limit Theorem.
- But it does *not* mean that $\overline{X}_n$ converges in distribution to a normal random variable.
- The Law of Large Numbers says that $\overline{X}_n$ converges almost surely (and in probability) to a constant, μ.
- So $\overline{X}_n$ converges to μ in distribution as well.

Why would we say that for large n, the sample mean is approximately $N(\mu, \frac{\sigma^2}{n})$?

Have $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1)$.

$$
\begin{aligned}
Pr\{\overline{X}_n \leq x\} &= Pr\left\{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} \\
&= Pr\left\{Z_n \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} \approx \Phi\left(\frac{\sqrt{n}(x-\mu)}{\sigma}\right)
\end{aligned}
$$

Suppose Y is *exactly* $N(\mu, \frac{\sigma^2}{n})$:

$$
\begin{aligned}
Pr\{Y \leq x\} &= Pr\left\{\frac{\sqrt{n}(Y - \mu)}{\sigma} \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} \\
&= Pr\left\{Z_n \leq \frac{\sqrt{n}(x-\mu)}{\sigma}\right\} = \Phi\left(\frac{\sqrt{n}(x-\mu)}{\sigma}\right)
\end{aligned}
$$

Convergence of random vectors I

1. Definitions (All quantities in boldface are vectors in $\mathbb{R}^m$ unless otherwise stated)
   - $\mathbf{T}_n \overset{a.s.}{\to} \mathbf{T}$ means $P\{\omega : \lim_{n\to\infty} \mathbf{T}_n(\omega) = \mathbf{T}(\omega)\} = 1$.
   - $\mathbf{T}_n \overset{P}{\to} \mathbf{T}$ means $\forall\epsilon > 0$, $\lim_{n\to\infty} P\{||\mathbf{T}_n - \mathbf{T}|| < \epsilon\} = 1$.
   - $\mathbf{T}_n \overset{d}{\to} \mathbf{T}$ means for every continuity point $\mathbf{t}$ of $F_{\mathbf{T}}$, $\lim_{n\to\infty} F_{\mathbf{T}_n}(\mathbf{t}) = F_{\mathbf{T}}(\mathbf{t})$.
2. $\mathbf{T}_n \overset{a.s.}{\to} \mathbf{T} \Rightarrow \mathbf{T}_n \overset{P}{\to} \mathbf{T} \Rightarrow \mathbf{T}_n \overset{d}{\to} \mathbf{T}$.
3. If $\mathbf{a}$ is a vector of constants, $\mathbf{T}_n \overset{d}{\to} \mathbf{a} \Rightarrow \mathbf{T}_n \overset{P}{\to} \mathbf{a}$.
4. Strong Law of Large Numbers (SLLN): Let $\mathbf{X}_1, \ldots \mathbf{X}_n$ be independent and identically distributed random vectors with finite first moment, and let $\mathbf{X}$ be a general random vector from the same distribution. Then $\overline{\mathbf{X}}_n \overset{a.s.}{\to} E(\mathbf{X})$.
5. Central Limit Theorem: Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors with expected value vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Then $\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})$ converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$.

Convergence of random vectors II

6. Slutsky Theorems for Convergence in Distribution:
   1. If $\mathbf{T}_n \in \mathbb{R}^m$, $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ and if $f : \mathbb{R}^m \to \mathbb{R}^q$ (where $q \leq m$) is continuous except possibly on a set C with $P(\mathbf{T} \in C) = 0$, then $f(\mathbf{T}_n) \xrightarrow{d} f(\mathbf{T})$.
   2. If $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ and $(\mathbf{T}_n - \mathbf{Y}_n) \xrightarrow{P} 0$, then $\mathbf{Y}_n \xrightarrow{d} \mathbf{T}$.
   3. If $\mathbf{T}_n \in \mathbb{R}^d$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{T}_n \xrightarrow{d} \mathbf{T}$ and $\mathbf{Y}_n \xrightarrow{P} \mathbf{c}$, then

$$\begin{pmatrix} \mathbf{T}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{T} \\ \mathbf{c} \end{pmatrix}$$

Convergence of random vectors III

7. Slutsky Theorems for Convergence in Probability:
   1. If $\mathbf{T}_n \in \mathbb{R}^m$, $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ and if $f : \mathbb{R}^m \to \mathbb{R}^q$ (where $q \leq m$) is continuous except possibly on a set C with $P(\mathbf{T} \in C) = 0$, then $f(\mathbf{T}_n) \xrightarrow{P} f(\mathbf{T})$.
   2. If $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ and $(\mathbf{T}_n - \mathbf{Y}_n) \xrightarrow{P} 0$, then $\mathbf{Y}_n \xrightarrow{P} \mathbf{T}$.
   3. If $\mathbf{T}_n \in \mathbb{R}^d$, $\mathbf{Y}_n \in \mathbb{R}^k$, $\mathbf{T}_n \xrightarrow{P} \mathbf{T}$ and $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$, then

$$\begin{pmatrix} \mathbf{T}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \mathbf{T} \\ \mathbf{Y} \end{pmatrix}$$

Convergence of random vectors IV

8. Delta Method (Theorem of Cramér, Ferguson p. 45): Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be such that the elements of $\dot{\mathrm{g}}(\mathbf{x}) = \left[\frac{\partial g_i}{\partial x_j}\right]_{k\times d}$ are continuous in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^d$. If $\mathbf{T}_n$ is a sequence of d-dimensional random vectors such that $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \stackrel{d}{\to} \mathbf{T}$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \stackrel{d}{\to} \dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{T}$. In particular, if $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \stackrel{d}{\to} \mathbf{T} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \stackrel{d}{\to} \mathbf{Y} \sim N(\mathbf{0}, \dot{\mathrm{g}}(\boldsymbol{\theta})\boldsymbol{\Sigma}\dot{\mathrm{g}}(\boldsymbol{\theta})')$.

An application of the Slutsky Theorems

- Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} ?(\mu, \sigma^2)$
- By CLT, $Y_n = \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} Y \sim N(0, \sigma^2)$
- Let $\widehat{\sigma}_n$ be *any* consistent estimator of σ.
- Then by 6.3, $\mathbf{T}_n = \begin{pmatrix} Y_n \\ \widehat{\sigma}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Y \\ \sigma \end{pmatrix} = \mathbf{T}$
- The function $f(x, y) = x/y$ is continuous except if $y = 0$ so by 6.1,

$$f(\mathbf{T}_n) = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\widehat{\sigma}_n} \xrightarrow{d} f(\mathbf{T}) = \frac{Y}{\sigma} \sim N(0, 1)$$

Univariate delta method

In the multivariate Delta Method 8, the matrix $\dot{\mathrm{g}}(\boldsymbol{\theta})$ is a Jacobian. The univariate version of the delta method says that if $\sqrt{n}\,(T_n - \theta) \stackrel{d}{\rightarrow} T$ and $g''(x)$ is continuous in a neighbourhood of θ, then

$$\sqrt{n}\,(g(T_n) - g(\theta)) \stackrel{d}{\rightarrow} g'(\theta)\,T.$$

When using the Central Limit Theorem, *especially* if there is a $\theta \neq \mu$ in the model, it's safer to write

$$\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \stackrel{d}{\rightarrow} g'(\mu)\,T.$$

and then substitute for μ in terms of θ.

Example: Geometric distribution

Let $X_1, \ldots, X_n$ be a random sample from a distribution,with probability mass function $p(x|\theta) = \theta(1-\theta)^{x-1}$ for $x = 1, 2, \ldots$, where $0 < \theta < 1$.

So, $E(X_i) = \frac{1}{\theta}$ and $Var(X_i) = \frac{1-\theta}{\theta^2}$.

The maximum likelihood estimator of θ is $\widehat{\theta} = \frac{1}{\overline{X}_n}$. Using the Central Limit Theorem and the delta method, find the approximate large-sample distribution of $\widehat{\theta}$.

Solution: Geometric distribution

$\mu = \frac{1}{\theta}$ and $\sigma^2 = \frac{1-\theta}{\theta^2}$

CLT says $\sqrt{n}\left(\overline{X}_n - \mu\right) \stackrel{d}{\rightarrow} T \sim N(0, \frac{1-\theta}{\theta^2})$

Delta method says $\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \stackrel{d}{\rightarrow} g'(\mu)\, T$.

$g(x) = \frac{1}{x} = x^{-1}$

$g'(x) = -x^{-2}$

So,

$$
\begin{aligned}
\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) &= \sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right) \\
&= \sqrt{n}\left(\widehat{\theta} - \theta\right) \\
&\stackrel{d}{\rightarrow} g'(\mu)\, T = -\frac{1}{\mu^2}\, T \\
&= -\theta^2\, T \sim N\left(0, \theta^4 \cdot \frac{1-\theta}{\theta^2}\right)
\end{aligned}
$$

Asymptotic distribution of $\widehat{\theta} = \frac{1}{\overline{X}_n}$

Approximate large-sample distribution

Have $Y_n = \sqrt{n}\left(\widehat{\theta} - \theta\right) \stackrel{\cdot}{\sim} N(0, \theta^2(1-\theta))$.

$$\text{So } \frac{Y_n}{\sqrt{n}} = \left(\widehat{\theta} - \theta\right) \stackrel{\cdot}{\sim} N\left(0, \frac{\theta^2(1-\theta)}{n}\right)$$
$$\text{And } \frac{Y_n}{\sqrt{n}} + \theta = \widehat{\theta} \stackrel{\cdot}{\sim} N\left(\theta, \frac{\theta^2(1-\theta)}{n}\right)$$

We'll say that $\widehat{\theta} = \frac{1}{\overline{X}_n}$ is approximately (or asymptotically) $N\left(\theta, \frac{\theta^2(1-\theta)}{n}\right)$.

Another example of $\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \stackrel{d}{\rightarrow} g'(\mu)\, T$

Don't lose your head

Let $X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} ?(\mu, \sigma^2)$

CLT says $\sqrt{n}(\overline{X}_n - \mu) \stackrel{d}{\rightarrow} T \sim N(0, \sigma^2)$

Let $g(x) = x^2$

Delta method says $\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \stackrel{d}{\rightarrow} g'(\mu)\, T$.

So $\sqrt{n}\left(\overline{X}_n^2 - \mu^2\right) \stackrel{d}{\rightarrow} 2\mu\, T \sim N(0, 4\mu^2\sigma^2)$

Really? What if $\mu = 0$?

If $\mu = 0$ then $\sqrt{n}\left(\overline{X}_n^2 - \mu^2\right) = \sqrt{n}\,\overline{X}_n^2 \stackrel{d}{\rightarrow} 2\mu T = 0$

$\Rightarrow \sqrt{n}\,\overline{X}_n^2 \stackrel{p}{\rightarrow} 0$.

Already know from continuous mapping that $\overline{X}_n^2 \stackrel{p}{\rightarrow} \mu^2 = 0$.

Delta method reveals *faster convergence*.

Also ...

Have $\sqrt{n}\,\overline{X}_n^2 \xrightarrow{p} 0$. If we add another $\sqrt{n}$ and if (say) $\sigma^2 = 1$ as well as $\mu = 0$,

$$n\overline{X}_n^2 = \left(\sqrt{n}(\overline{X}_n - \mu)\right)^2 \xrightarrow{d} Z^2 \sim \chi^2(1)$$

If $\sigma^2 \neq 1$, the target is Gamma($\alpha = \frac{1}{2}$, $\beta = 2\sigma$)

The delta method comes from Taylor's Theorem

Taylor's Theorem: Let the nth derivative $f^{(n)}$ be continuous in $[a, b]$ and differentiable in (a, b), with x and x_0 in (a, b). Then there exists a point ξ between x and x_0 such that

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)(x - x_0)^2}{2!} + \ldots \\ &+ \frac{f^{(n)}(x_0)(x - x_0)^n}{n!} + \frac{f^{(n+1)}(\xi)(x - x_0)^{n+1}}{(n+1)!} \end{aligned}$$

where $R_n = \frac{f^{(n+1)}(\xi)(x - x_0)^{n+1}}{(n+1)!}$ is called the *remainder term*.
If $R_n \to 0$ as $n \to \infty$, the resulting infinite series is called the *Taylor Series* for $f(x)$.

Taylor's Theorem with two terms plus remainder

Very common in applications

Let $g(x)$ be a function for which $g''(x)$ is continuous in an open interval containing $x = \theta$. Then

$$g(x) = g(\theta) + g'(\theta)(x - \theta) + \frac{g''(\theta^*)(x-\theta)^2}{2!}$$

where θ^* is between x and θ.

Delta method

Using $g(x) = g(\theta) + g'(\theta)(x-\theta) + \frac{1}{2}g''(\theta^*)(x-\theta)^2$

Let $\sqrt{n}(T_n - \theta) \xrightarrow{d} T$ so that $T_n \xrightarrow{p} \theta$.

$$\begin{aligned}
\sqrt{n}\left(g(T_n) - g(\theta)\right) &= \sqrt{n}\left(g(\theta) + g'(\theta)(T_n-\theta) + \frac{1}{2}g''(\theta_n^*)(T_n-\theta)^2 - g(\theta)\right) \\
&= \sqrt{n}\left(g'(\theta)(T_n-\theta) + \frac{1}{2}g''(\theta_n^*)(T_n-\theta)^2\right) \\
&= g'(\theta)\sqrt{n}(T_n-\theta) \\
&\quad + \frac{1}{2}g''(\theta_n^*)\cdot\sqrt{n}(T_n-\theta)\cdot(T_n-\theta) \\
&\xrightarrow{d} g'(\theta)T + 0
\end{aligned}$$

A variance-stabilizing transformation

An application of the delta method

- Because the Poisson process is such a good model, count data often have approximate Poisson distributions.
- Let $X_1, \ldots, X_n \stackrel{i.i.d}{\sim}$ Poisson(λ)
- $E(X_i) = Var(X_i) = \lambda$
- $Z_n = \frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\overline{X}_n}} \stackrel{d}{\rightarrow} Z \sim N(0,1)$
- Could say $\overline{X}_n \stackrel{.}{\sim} N(\lambda, \lambda/n)$ and $\sum_{i=1}^n X_i \stackrel{.}{\sim} N(n\lambda, \lambda)$.
- Because the sum of independent Poissons is Poisson, this means Poisson-distributed variables with large λ are approximately normal.
- For analysis with normal linear models, approximate normality is good. Variance that depends on $E(Y_i)$ is not good.
- Can we fix it?

Variance-stabilizing transformation continued

- CLT says $\sqrt{n}(\overline{X}_n - \lambda) \stackrel{d}{\rightarrow} T \sim N(0, \lambda)$.
- Delta method says
  $\sqrt{n}\left(g(\overline{X}_n) - g(\lambda)\right) \stackrel{d}{\rightarrow} g'(\lambda)\, T = Y \sim N\left(0, g'(\lambda)^2\, \lambda\right)$
- If $g'(\lambda) = \frac{1}{\sqrt{\lambda}}$, then $Y \sim N(0, 1)$.

An elementary differential equation: $g'(x) = \frac{1}{\sqrt{x}}$

Solve by separation of variables

$$\begin{aligned} & \frac{dg}{dx} = x^{-1/2} \\ \Rightarrow \;\; & dg = x^{-1/2}\,dx \\ \Rightarrow \;\; & \int dg = \int x^{-1/2}\,dx \\ \Rightarrow \;\; & g(x) = \frac{x^{1/2}}{1/2} + c = 2x^{1/2} + c \end{aligned}$$

We have found

$$\begin{aligned}\sqrt{n}\left(g(\overline{X}_n) - g(\lambda)\right) &= \sqrt{n}\left(2\overline{X}_n^{1/2} - 2\lambda^{1/2}\right) \\ &\stackrel{d}{\to} Z \sim N(0,1)\end{aligned}$$

So,

- We could say that $\sqrt{\overline{X}_n}$ is asymptotically normal, with (asymptotic) mean $\sqrt{\lambda}$ and (asymptotic) variance $\frac{1}{4n}$.
- This calculation could justify a square root transformation for count data.
- Note that the transformation is increasing, so if Y_i is number of visitors to a website, $\sqrt{Y_i}$ could still be called "popularity."

The arcsin-square root transformation
For proportions

Sometimes, variable values consist of proportions, one for each case.

- For example, cases could be high schools.
- The variable of interest is the proportion of students who enroll in university the year after graduation.
- This is an example of *aggregated data*.

The advice you sometimes get

Still

When a proportion is the response variable in a regression, use the *arcsin square root* transformation.

That is, if the proportions are $P_1, \ldots, P_n$, let

$$Y_i = \sin^{-1}(\sqrt{P_i})$$

and use the Y_i values in your regression.

Why?

It's a variance-stabilizing transformation (details omitted).

That was fun, but it was all univariate.

Because

- The multivariate CLT establishes convergence to a multivariate normal, and
- Vectors of MLEs are approximately multivariate normal for large samples, and
- The multivariate delta method can yield the asymptotic distribution of useful functions of the MLE vector,

We need to look at random vectors and the multivariate normal distribution.

Copyright Information