Large sample tools[1]

STA442/2101 Fall 2013

[1] See last slide for copyright information.

Background Reading: Davison's *Statistical models*

- For completeness, look at Section 2.1, which presents some basic applied statistics in an advanced way.
- Especially 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, $\omega \in \Omega$

- Observe whether a single individual is male or female: $\Omega = \{F, M\}$
- Pair of individuals; observe their genders in order: $\Omega = \{(F, F), (F, M), (M, F), (M, M)\}$
- Select n people and count the number of females: $\Omega = \{0, \ldots, n\}$

For limits problems, the points in Ω are infinite sequences.

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 $T_n \stackrel{a.s.}{\rightarrow} T$ if

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

- 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.

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
- Power
- Weather model
- Performance of statistical methods
- Need confidence intervals for estimated probabilities.

Estimating power by simulation

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?

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

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

```
# 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 $E[g(X)]$

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

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 $\mu = E[g(X)]$ is routine,

Convergence in Probability

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

if for all $\epsilon > 0$,

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

Convergence in probability (say to a constant θ) means no matter how small the interval around θ, for large enough n (that is, for all $n > N_1$) the probability of getting that close to θ is as close to one as you like.

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 $T_n \stackrel{P}{\rightarrow} \theta$.

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

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.

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

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

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

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

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,

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

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

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

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)
   - $\mathbf{T}_n \stackrel{a.s.}{\rightarrow} \mathbf{T}$ means $P\{\omega : \lim_{n\to\infty} \mathbf{T}_n(\omega) = \mathbf{T}(\omega)\} = 1$.
   - $\mathbf{T}_n \stackrel{P}{\rightarrow} \mathbf{T}$ means $\forall\epsilon > 0$, $\lim_{n\to\infty} P\{||\mathbf{T}_n - \mathbf{T}|| < \epsilon\} = 1$.
   - $\mathbf{T}_n \stackrel{d}{\rightarrow} \mathbf{T}$ 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. $\mathbf{T}_n \stackrel{a.s.}{\rightarrow} \mathbf{T} \Rightarrow \mathbf{T}_n \stackrel{P}{\rightarrow} \mathbf{T} \Rightarrow \mathbf{T}_n \stackrel{d}{\rightarrow} \mathbf{T}$.
3. If $\mathbf{a}$ is a vector of constants, $\mathbf{T}_n \stackrel{d}{\rightarrow} \mathbf{a} \Rightarrow \mathbf{T}_n \stackrel{P}{\rightarrow} \mathbf{a}$.
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 $\overline{\mathbf{X}}_n \stackrel{a.s.}{\rightarrow} E(\mathbf{X})$.
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 $\boldsymbol{\Sigma}$. Then $\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu})$ converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

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 $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T}$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \dot{\mathrm{g}}(\boldsymbol{\theta})\mathbf{T}$. In particular, if $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \dot{\mathrm{g}}(\boldsymbol{\theta})\boldsymbol{\Sigma}\dot{\mathrm{g}}(\boldsymbol{\theta})')$.

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

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

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

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

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

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

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

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

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

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

Don't lose your head

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

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

Let $g(x) = x^2$

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

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

Really? What if $\mu = 0$?

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

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

Faster convergence.

On the other hand . . .

Have $\sqrt{n}\,\overline{X}_n^2 \stackrel{p}{\rightarrow} 0$, but if (say) $\sigma^2 = 1$,

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

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

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 $X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{Poisson}(\lambda)$
- $E(X_i) = Var(X_i) = \lambda$
- $Z_n = \frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\overline{X}_n}} \overset{d}{\rightarrow} Z \sim N(0,1)$
- An approximate large-sample confidence interval for λ is

$$\overline{X}_n \pm z_{\alpha/2} \sqrt{\frac{\overline{X}_n}{n}}$$

- Can we do better?

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

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

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.
- How about a better confidence interval for λ?

Seeking a better confidence interval for λ

$$
\begin{aligned}
1-\alpha &= \Pr\{-z_{\alpha/2} < Z < z_{\alpha/2}\} \\
&\approx \Pr\{-z_{\alpha/2} < 2\sqrt{n}\left(\overline{X}_n^{1/2} - \lambda^{1/2}\right) < z_{\alpha/2}\} \\
&= \Pr\left\{\sqrt{\overline{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}} < \sqrt{\lambda} < \sqrt{\overline{X}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right\} \\
&= \Pr\left\{\left(\sqrt{\overline{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2 < \lambda < \left(\sqrt{\overline{X}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2\right\},
\end{aligned}
$$

where the last equality is valid provided $\sqrt{\overline{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}} \geq 0$.

Compare the confidence intervals

Variance-stabilized CI is

$$
\begin{aligned}
& \left(\left(\sqrt{\overline{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2 , \left(\sqrt{\overline{X}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2 \right) \\
= & \left(\overline{X}_n - 2\sqrt{\overline{X}_n}\frac{z_{\alpha/2}}{2\sqrt{n}} + \frac{z_{\alpha/2}^2}{4n} , \overline{X}_n + 2\sqrt{\overline{X}_n}\frac{z_{\alpha/2}}{2\sqrt{n}} + \frac{z_{\alpha/2}^2}{4n} \right) \\
= & \left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}} + \frac{z_{\alpha/2}^2}{4n} , \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}} + \frac{z_{\alpha/2}^2}{4n} \right)
\end{aligned}
$$

Compare to the ordinary (Wald) CI

$$
\left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}} , \overline{X}_n + z_{\alpha/2}\sqrt{\frac{\overline{X}_n}{n}} \right)
$$

Variance-stabilized CI is just like the ordinary CI

Except shifted to the right by $\frac{z^2_{\alpha/2}}{4n}$.

- If there is a difference in performance, we will see it for small n.
- Try some simulations.
- Is the coverage probability closer?

Try $n = 10$, True $\lambda = 1$

Illustrate the code first

```
> # Variance stabilized Poisson CI
> n = 10; lambda=1; m=10; alpha = 0.05; set.seed(9999)
> z = qnorm(1-alpha/2)
> cover1 = cover2 = NULL
> for(sim in 1:m)
+     {
+     x = rpois(n,lambda); xbar = mean(x); xbar
+     a1 = xbar - z*sqrt(xbar/n); b1 = xbar + z*sqrt(xbar/n)
+     shift = z^2/(4*n)
+     a2 = a1+shift; b2 = b1+shift
+     cover1 = c(cover1,(a1 < lambda && lambda < b1))
+     cover2 = c(cover2,(a2 < lambda && lambda < b2))
+     } # Next sim
> rbind(cover1,cover2)
       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
cover1 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
cover2 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
> mean(cover1)
[1] 0.9
```

Code for Monte Carlo sample size = 10,000 simulations

```
# Now the real simulation
n = 10; lambda=1; m=10000; alpha = 0.05; set.seed(9999)
z = qnorm(1-alpha/2)
cover1 = cover2 = NULL
for(sim in 1:m)
    {
    x = rpois(n,lambda); xbar = mean(x); xbar
    a1 = xbar - z*sqrt(xbar/n); b1 = xbar + z*sqrt(xbar/n)
    shift = z^2/(4*n)
    a2 = a1+shift; b2 = b1+shift
    cover1 = c(cover1,(a1 < lambda && lambda < b1))
    cover2 = c(cover2,(a2 < lambda && lambda < b2))
    } # Next sim
p1 = mean(cover1); p2 = mean(cover2)
# 99 percent margins of error
me1 = qnorm(0.995)*sqrt(p1*(1-p1)/m); me1 = round(me1,3)
me2 = qnorm(0.995)*sqrt(p1*(1-p1)/m); me2 = round(me2,3)
cat("Coverage of ordinary CI = ",p1,"plus or minus ",me1,"\n")
cat("Coverage of variance-stabilized CI = ",p2,
"plus or minus ",me2,"\n")
```

Results for $n = 10$, $\lambda = 1$ and 10,000 simulations

```
Coverage of ordinary CI =  0.9292 plus or minus  0.007

Coverage of variance-stabilized CI =  0.9556 plus or minus  0.007

> # Does CI include 0.95?
> # Look at estimate (too high) minus margin of error.
> p2-me2
[1] 0.9486
```

Results for $n = 100$

$\lambda = 1$ and 10,000 simulations

```
Coverage of ordinary CI =  0.9448 plus or minus  0.006

Coverage of variance-stabilized CI =  0.9473 plus or minus  0.006

> p1+me1
[1] 0.9508
```

The arcsin-square root transformation

For proportions

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

- For example, cases could be hospitals.
- The variable of interest is the proportion of patients who came down with something *unrelated* to their reason for admission – hospital-acquired infection.
- This is an example of *aggregated data*.

The advice you often get

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.

- The proportions are little sample means: $P_i = \frac{1}{m}\sum_{j=1}^m X_{i,j}$
- Drop the i for now.
- $X_1, \ldots, X_m$ may not be independent, but let's pretend.
- $P = \overline{X}_m$
- Approximately, $\overline{X}_m \sim N\left(\theta, \frac{\theta(1-\theta)}{m}\right)$
- Normality is good.
- Variance that depends on the mean θ is not so good.

Apply the delta method

Central Limit Theorem says

$$\sqrt{m}(\overline{X}_m - \theta) \xrightarrow{d} T \sim N\left(0, \theta(1-\theta)\right)$$

Delta method says

$$\sqrt{m}\left(g(\overline{X}_m) - g(\theta)\right) \xrightarrow{d} Y \sim N\left(0, g'(\theta)^2\, \theta(1-\theta)\right).$$

Want a function $g(x)$ with

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

Try $g(x) = \sin^{-1}(\sqrt{x})$.

Chain rule to get $\frac{d}{dx}\sin^{-1}\left(\sqrt{x}\right)$

"Recall" that $\frac{d}{dx}\sin^{-1}(x) = \frac{1}{\sqrt{1-x^2}}$. Then,

$$\begin{aligned}\frac{d}{dx}\sin^{-1}\left(\sqrt{x}\right) &= \frac{1}{\sqrt{1-\sqrt{x}^2}} \cdot \frac{1}{2}x^{-1/2} \\ &= \frac{1}{2\sqrt{x(1-x)}}.\end{aligned}$$

Conclusion:

$$\sqrt{m}\left(\sin^{-1}\left(\sqrt{\overline{X}_m}\right) - \sin^{-1}\left(\sqrt{\theta}\right)\right) \stackrel{d}{\to} Y \sim N\left(0, \frac{1}{4}\right)$$

So the arcsin-square root transformation stabilizes the variance

- $Y \sim N\left(0, \frac{1}{4}\right)$ means the variance no longer depends on the probability that the proportion is estimating.
- Does not quite *standardize* the proportion, but that's okay for regression.
- Potentially useful for non-aggregated data too.
- If we want to do a regression on aggregated data, the point we have reached is that approximately,

$$Y_i \sim N\left(\sin^{-1}\left(\sqrt{\theta_i}\right), \frac{1}{4m_i}\right)$$

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