

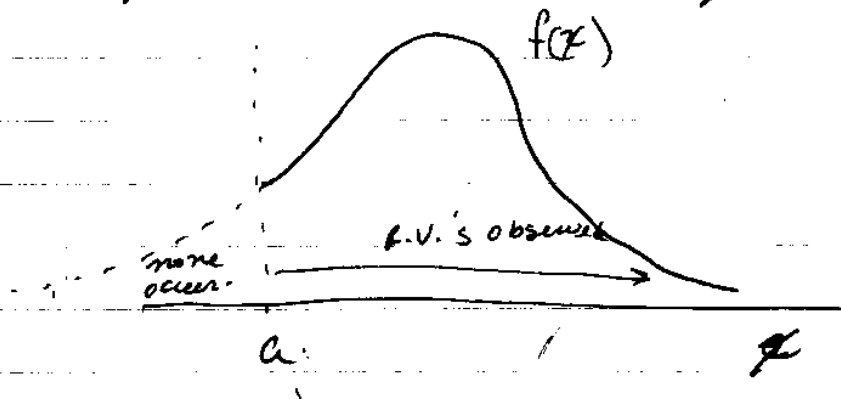
## Truncation and censoring

Theorem: Density of a truncated R.V.

If a continuous R.V.,  $X$ , has p.d.f.  $f(x)$  and  $a$  is a constant

$$f(x|X > a) = \frac{f(x)}{\text{Prob}(X > a)}$$

This amounts to the fact that if  $f(x)$  is truncated, then we need to renormalize the p.d.f. so that it integrates to 1



Example:  $X \sim U(0,1)$      $f(x) = 1$      $0 \leq x \leq 1$   
Let  $a = 1/3$

$$f(x|X > 1/3) = \frac{f(x)}{\text{Prob}(X > 1/3)} = \frac{1}{2/3} = 3/2$$

$$f(x) = 3/2 \quad 1/3 < x \leq 1$$

Truncated normal

If  $x \sim N(\mu, \sigma^2)$  then

normalize ( $N(0,1)$ )

$$\begin{aligned} \text{Prob}(x > a) &= 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \\ &= 1 - \Phi(\alpha) \end{aligned}$$

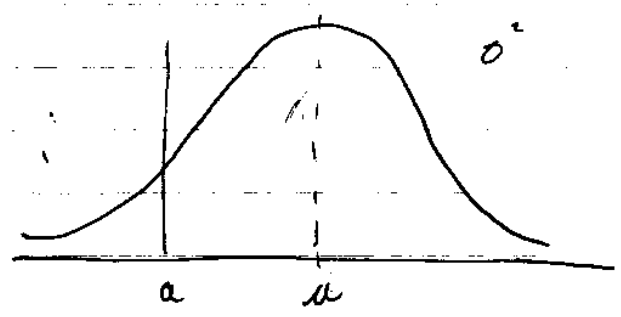
where  $\alpha = \frac{a-\mu}{\sigma}$

So,  $f(x|x > a) = \frac{f(x)}{1 - \Phi(\alpha)}$

$$= \frac{(2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{1 - \Phi(\alpha)}$$

$$= \frac{1/\sigma \phi\left(\frac{x-\mu}{\sigma}\right)}{1 - \Phi(\alpha)}$$

where  $\phi$  is  $\frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$  i.e. std normal p.d.f.



$x \sim N(\mu, \sigma^2)$   
 $z \sim N(0,1)$  where  $z = \frac{x-\mu}{\sigma}$   
 at  $x=a =$

Moments of a Truncated Dist.

$$E[X|X > a] = \int_a^{\infty} x f(x|x > a) dx$$

Example:  $E[X|X > \frac{1}{3}]$   $X \sim U(0,1)$   
 $\mu = \frac{1}{3}$

$$\int_{\frac{1}{3}}^1 x^{\frac{3}{2}} dx = \frac{\frac{3}{2} \cdot x^{\frac{5}{2}}}{\frac{5}{2}} \Big|_{\frac{1}{3}}^1 = \frac{5 \cdot 2}{4} - \frac{3}{36} = \frac{24}{36} = \frac{2}{3}$$

$$E(X|X > \frac{1}{3}) = \frac{2}{3}$$

$$\text{Var}(X|X > \frac{1}{3}) = \frac{1}{27} \quad \text{by virtue}$$

$$\text{Var} \left[ \overset{\text{of}}{\mathcal{U}(b,c)} \right] = \frac{(c-b)^2}{12}$$

$$\left(1 - \frac{1}{3}\right)^2 / 12$$

untruncated

$$X \sim \mathcal{U}(0,1)$$

$$E(X) = \frac{1}{2}$$

$$\text{Var}(X) = \frac{1}{12}$$

$$\frac{\frac{2}{3} \cdot \frac{4}{9}}{9}$$

$$\frac{4}{9} \cdot \frac{1}{14.5} = \frac{2}{14.5}$$

Truncated (from below)

$$E(X|X > \frac{1}{3}) = \frac{2}{3} > \frac{1}{2}$$

$$\text{Var}(X|X > \frac{1}{3}) = \frac{1}{27} < \frac{1}{12}$$

mean bigger  
var smaller

## Moments of Trunc Normal

If  $X \sim N(\mu, \sigma^2)$  and  $a$  is const

$$\begin{aligned} E[X | \text{truncation}] &= \mu + \sigma \lambda(\alpha) \\ \text{Var}[X | \text{ " "}] &= \sigma^2 (1 - \xi(\alpha)) \end{aligned} \quad (2)$$

where  $\alpha = \frac{a - \mu}{\sigma}$

Hazard Function  $\rightarrow \lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$  if  $x > a$

$\lambda(\alpha) = \frac{-\phi(\alpha)}{\Phi(\alpha)}$  if  $x < a$

$$\xi(\alpha) = \lambda(\alpha) [\lambda(\alpha) - \alpha]$$

Note:  $0 < \xi(\alpha) < 1$  since variance of trunc will be smaller see (2) above.

$$\frac{d\phi(\alpha)}{d\alpha} = -\alpha \phi(\alpha) \quad \text{same as in probab.}$$

$\lambda(\alpha)$  is called Inverse Mills Ratio

## Truncated Regression Model

$$\text{let } u = \gamma_i' \beta$$

$$y_i = \gamma_i' \beta + e_i \quad e_i \sim N(0, \sigma^2)$$

$$y_i | \gamma_i \sim N(\gamma_i' \beta, \sigma^2)$$

Truncation pt.

$$E[y_i | y_i > a] = \gamma_i' \beta + \sigma \frac{\phi\left(\frac{a - \gamma_i' \beta}{\sigma}\right)}{1 - \Phi\left[\frac{a - \gamma_i' \beta}{\sigma}\right]}$$

conditional mean is a nonlinear function of  $\beta, \gamma$ .

Consider the subpopulation  $y_i > a$ .

$$E[y_i | y_i > a] = \gamma_i' \beta + \sigma \lambda(\alpha_i)$$

$$\alpha_i = \frac{a - \gamma_i' \beta}{\sigma}$$

$$\frac{\partial E[\gamma_i | y_i > a]}{\partial \gamma} = \beta + \sigma \frac{\partial \lambda_i}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial \gamma}$$

$$= \beta + \sigma (\lambda_i^2 - \alpha_i \lambda_i) \left(-\frac{\beta}{\sigma}\right)$$

$$= \beta (1 - \lambda_i^2 + \alpha_i \lambda_i)$$

$$= \beta (1 - \delta(\alpha_i))$$

note marginal effects in sub popl  
will understate marginal effect in  
total popl.

$$\text{Variance } (y_i | y_i > a) = \sigma^2 (1 - \Phi(d_i))$$

Heteroscedastic.

Pop'l inferences, OLS is biased, inconsistent and heteroscedastic.

MLE

$$f(y_i) = \frac{\frac{1}{\sigma} \phi\left(\frac{y_i - \pi_i' \beta}{\sigma}\right)}{1 - \Phi\left(\frac{y_i - \pi_i' \beta}{\sigma}\right)}$$

$$L = \ln(l) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \pi_i' \beta)^2 - \sum \ln\left(1 - \Phi\left(\frac{a - \pi_i' \beta}{\sigma}\right)\right)$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^T \left( \frac{y_i - \pi_i' \beta}{\sigma^2} - \frac{\lambda_i}{\sigma} \right) \pi_i = 0$$

$$\frac{\partial L}{\partial \sigma^2} = \sum_{i=1}^T \left( -\frac{1}{2\sigma^2} + \frac{(y_i - \pi_i' \beta)^2}{2\sigma^4} - \frac{d_i \lambda_i}{\sigma^2} \right) = 0$$

$$d_i = \frac{a - \pi_i' \beta}{\sigma} \qquad \lambda_i = \frac{\phi(d_i)}{1 - \Phi(d_i)}$$

## Censored DATA

When a dependent variable is censored, values in a certain range are all transformed to the same value.

Example: Tickets demanded for events at Gallagher-Iba. Our only measure is # sold. When an event sells out we know that demand is larger than # sold. The # of tickets demanded is censored when transformed into # sold.

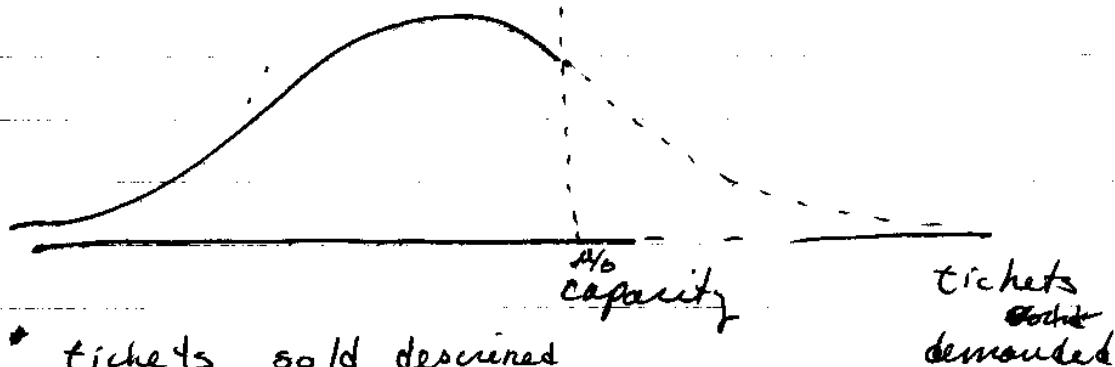
### Used Elsewhere

1. Household purchases of Durable goods
2. # extramarital affairs
3. # hours worked by women in LF
4. # arrests after release from prison

### Censored Normal Dist

similar in spirit to Truncated Normal and as before most theory is based on ~~non~~ censored normal (although others could be imagined)

In censored model The dist that applies to the sample is a mixture of discrete & continuous distributions



- \* tickets sold described by censored dist.
- \* tickets demanded by the full dist.

$$y = 0 \quad \text{if } y^* \leq 0$$

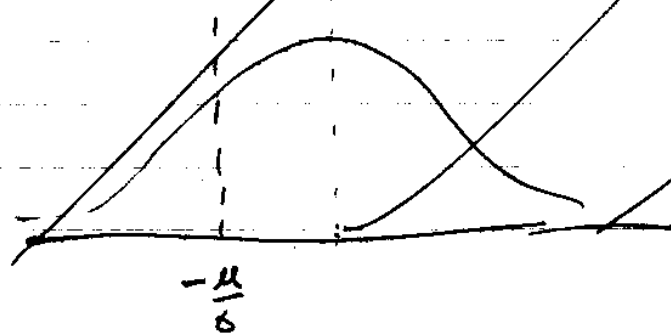
$$y = y^* \quad \text{if } y^* > 0$$

So, if  $y^* \sim N(\mu, \sigma^2)$

$$\text{Prob}(y=0) = \text{Prob}(y^* \leq 0) = \Phi\left(-\frac{\mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu}{\sigma}\right)$$

$$\text{Prob}(y^* > 0) = \text{describe } y^* \sim N(\mu, \sigma^2)$$

(censored above)



## Moments of Censored Normal

If  $y^* \sim N(\mu, \sigma^2)$  and  $y = a$  if  $y^* \leq a$   
 else  $y^* \quad y = y^*$

$$y = \begin{cases} a & y^* \leq a \\ y^* & y^* > a \end{cases}$$

Then

$$E[y] = \Phi \cdot a + (1 - \Phi)(\mu + \sigma\lambda)$$

and

$$\text{Var}(y) = \sigma^2(1 - \Phi) \left[ 1 - \delta + (\alpha - \lambda)^2 \Phi \right]$$

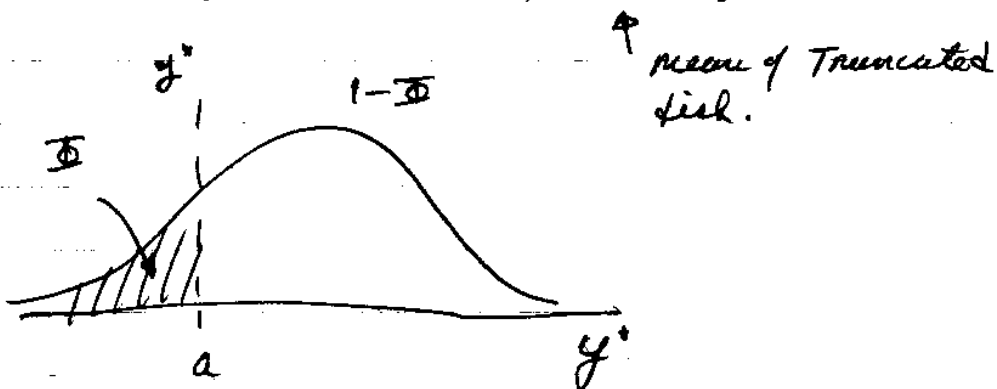
where

$$\Phi \equiv \Phi\left[\frac{a - \mu}{\sigma}\right] \equiv \Phi(\alpha) = \text{Prob}(y^* \leq a)$$

$$\lambda = \phi(\alpha) / (1 - \Phi) \quad \text{from Below}$$

$$\delta = \lambda^2 - \lambda\alpha$$

Proof:  $E(y) = P(y=a) \cdot E(y|y=a) + P(y>a) \cdot E(y|y>a)$   
 $= P(y^* \leq a) \cdot a + P(y^* > a) \cdot E(y^* | y^* > a)$   
 $= \Phi \cdot a + (1 - \Phi)(\mu + \sigma\lambda)$



For Trunc Above  
 redefine  $\lambda = \frac{-\phi}{\Phi}$  and change  $\Phi$  to  $1 - \Phi$

Variance: use fact

$$\text{Var}(y) = E(\text{cond var}) + \text{Var}(\text{cond mean})$$

Example:

Arena has 20,000 seats and sells out 25% of the time. If Avg Attendance is 18,000 (including sellouts) what is mean & std dev of demand for seats.

$$E(\text{sales}) = 20,000(1 - \Phi) + (\mu + \sigma\lambda)\Phi \quad (1)$$

considering from above, reverse also.

inverse  
 Mills  
 ratio

$$\lambda = \frac{-\phi(\alpha)}{\Phi(\alpha)} \quad \alpha = \frac{20,000 - \mu}{\sigma} \quad (2)$$

if 25% sellout  $\Phi = .75$ . Inverting this at  
 i.e.  $\Phi(.675) = .75$  .75 yield  $\alpha = .675$

$$\alpha = .675 \Rightarrow \lambda = \frac{-\phi(.675)}{.75} = \lambda = -.424$$

Est Avg Atten

$$18,000 = .25(20,000) + .75(\mu - .424\sigma) \quad (1)$$

$$\sigma(.675) = 20,000 - \mu \quad (2)$$

Solve 2 eq in 2 unk.

$$\sigma = 2426$$

$$\mu = 18362$$

If you assumed 18,000 applies to only non sellouts (truncated)

Since we exclude all obs for 18,000. Some of these would have been >.

$$18,000 = \mu - .4246$$

$$.6756 = 20,000 - \mu.$$

$$\sigma = 1820$$

$$\mu = 18,772$$

### Tobit Regression

$$y_i^* = \gamma_i' \beta + \epsilon_i$$

$$y_i = 0 \quad \text{if } y_i^* \leq 0$$

$$y_i = y_i^* \quad \text{if } y_i^* > 0$$

3 potential conditional mean functions to consider.

$$E(y_i^*) = \gamma_i' \beta$$

Not that useful if data censored always

NOTE: Assume  $\sigma = 1$

$$E[y_i | \gamma_i] = \Phi\left(\frac{\gamma_i' \beta}{\sigma}\right) (\gamma_i' \beta + \sigma \lambda_i)$$

$$\text{where } \lambda_i = \phi\left(\frac{\gamma_i' \beta}{\sigma}\right) / \Phi\left(\frac{\gamma_i' \beta}{\sigma}\right)$$

This applies to a random obs drawn from people that may or may not be censored.

If you only want info on uncensored ones,

$$E(y_i | x_i, y_i > 0) = \frac{x_i'(\beta + \sigma\lambda) \phi\left(\frac{x_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)}$$

From Truncated Model.

Which you use depends on purpose

- ✓ Predict tickets sold for coming event  
censored mean (second case)
- ✓ Need for new facility E(y<sub>i</sub>) (first case)

## Marginal Effects

$$\frac{\partial E(y_i | x_i)}{\partial x_i} = \beta$$

$$\frac{\partial E(y_i | x_i)}{\partial x_i} = \beta \Phi\left(\frac{x_i' \beta}{\sigma}\right) \quad \text{given censoring}$$

or more generally  $\frac{\partial E(y_i | x_i)}{\partial x_i} = \beta \cdot \text{Prob}(a < y_i < b)$   
Estimation for censoring above  $b$  and below  $a$ .

$$L = \ln L = \sum_{y_i > 0} -\frac{1}{2} \left[ \ln(2\pi) + \ln(\sigma^2) + \frac{(y_i - x_i' \beta)^2}{\sigma^2} \right] \\ + \sum_{y_i = 0} \ln \left[ 1 - \Phi\left(\frac{x_i' \beta}{\sigma}\right) \right]$$

$$\frac{\partial L}{\partial \beta} = -\frac{1}{\sigma} \sum_{y_i > 0} \frac{\phi_i x_i}{1 - \Phi_i} + \frac{1}{\sigma^2} \sum_{y_i > 0} (y_i - x_i' \beta) x_i$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{1}{2\sigma^3} \sum_{y_i > 0} \frac{(x_i' \beta) \phi_i}{1 - \Phi_i} - \frac{T_1}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{y_i > 0} (y_i - x_i' \beta)^2$$

$T_1 = \#$  of continuous obs.

$$f_i = \Phi_i \quad F_i = \Phi_i \quad z_i = \frac{f_i \beta}{\sigma}$$

$$I(\theta) = \begin{bmatrix} \sum_i a_i f_i' & \sum_i b_i f_i \\ \sum_i b_i f_i' & \sum_i c_i \end{bmatrix}$$

$$a_i = -\frac{1}{\sigma^2} \left( z_i f_i - \frac{f_i^2}{1-F_i} - F_i \right)$$

$$b_i = \frac{1}{2\sigma^3} \left( z_i^2 f_i + f_i - \frac{z_i f_i^2}{1-F_i} \right)$$

$$c_i = -\frac{1}{4\sigma^4} \left( z_i^3 f_i + z_i f_i - \frac{z_i^2 f_i^2}{1-F_i} - 2F_i \right)$$

Solve using NR, MOS, BHHH. Let  $\theta = \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}$

$$\sqrt{T} (\hat{\theta}_{MLE} - \theta) \xrightarrow{d} N(0, \lim_{T \rightarrow \infty} [I(\theta)]^{-1})$$

etc.