- We want to represent counts
  Poisson R.V.'s capture the idea of uniformity.

  - The count in an interval
    - depends on length of interval
    - in fact, $\propto$ length
  - Events are independent.

  (Idea works in multiple dimensions).

  If $X$ is Poisson, intensity $\lambda$.

  $E(X) = \lambda$

  $\text{Var}(X) = \lambda$

  Notice $\lambda$ has units (e.g. $\frac{\#}{s}$, etc.

  and depends on scale of interval.
PMF:

\[ P(X = n \mid \text{unit \, \text{rate} } \lambda) = \frac{e^{-\lambda} \lambda^n}{n!} \]

Straightforward series manipulation gives

- This is a PMF
- Expectation
- Variance

Now assume whatever we're watching is Poisson

- Types are independent

These assumptions are a stretch; words aren't like this; neither are animals or objects; but they're simple + generic
Notice:

1) If I observe a Poisson RV for an interval longer than 1

\[ P(X = n \mid \text{interval length } t \lambda) = \frac{e^{-\lambda t} (\lambda t)^n}{n!} \]

2) If I observe Poisson RV for \( N \) unit intervals, seeing count \( n_i \)
   in the \( i \)'th, Max likelihood est \( \lambda \) is

\[ \lambda^* = \frac{1}{N} \sum n_i \]

3) If I observe poisson RV for \( N \) unit intervals, \( i \)th has length \( \lambda_i \),
   and see \( n_i \) in \( i \)'th, Max likelihood gives

\[ \lambda^* = \frac{1}{N} \sum_i \left( \frac{n_i}{\lambda_i} \right) \]
Now assume we see \{ words, objects \} for an interval (which could be time or space). We choose a scale so this interval is $[-1, \infty]$. Each word type has intensity

$$\lambda_{\omega} = \text{intensity of type } \omega$$

which is unknown. Natural to try and draw conclusions from word type counts,

$$n_{\omega c} = \left[ \frac{\text{number of word types}}{\text{that appear } x \text{ times}} \right]$$

Natural because Max likelihood on individual words is no help.
Also, assume future is like the past.

i.e. if a word has \( \lambda_0 \) in \([1, 0]\)

it has this \( \lambda \) later.

E+T phrase this as

"conditionally binomial"
Mk est. $\theta \propto \lambda$ for words we haven't seen is 0.

But $n_1$ large compared to $n_2$, etc. suggests there are words types where

- we haven't seen them
- $\lambda_w > 0$

So we should be looking at $G(\lambda)$.

$$G(\lambda) = P(\lambda_w \leq \lambda)$$

- clearly, a discrete distribution
- represents CDF (cumulative dist. function)

$$dG(\lambda) = p(\lambda_w = \lambda) d\lambda$$

S functions or atoms
Now

\[ p(\text{a word type has count } x | \lambda_w) \]

\[ = e^{-\lambda_w} \frac{\lambda_w^x}{x!} \]

\[ \therefore \quad \mathbb{E}[\text{# of word types w/ count } x] \]

\[ = \sum_{\text{all word types}} p(\text{word type } i \text{ has count } x) \]

\[ = C \int e^{-\lambda} \frac{\lambda^x}{x!} p(G(\lambda)) \, d\lambda = \eta_x \]

\[ \text{total # of word types, unknown!} \]
Another way to look at this is that

\[ d\Gamma(\lambda) = C dG(\lambda) \] is a measure

(like a PDF, i.e., but doesn't \( \int \) to 1)

Notice that \( C \) could be hard to get, because we could have support for \( G(\lambda) \) at, say,

\[ \lambda = 10^{-12} \]

there is a word we see about once in \( 10^{12} \) intervals.

affects \( C \), but not a significant effect on observations.
Now \( \eta_x \) is the observed value of an RV call it \( \eta_x \)

we have

\[ E(\eta_x) = \eta_x \]

reasonable approx

\[ \text{Var}(\eta_x) = \eta_x \]

\( \sum \) of \( \eta_x \)

\( \eta_x \sim N(\eta_x, \sqrt{\eta_x}) \)

\( \sum \) of random variables

This will come in useful.
Now consider

\[ \Delta(t) = E[\text{\# \& types seen in } [0, t], \text{ but not in } [-1, 0]] \]

then

\[ \Delta(t) = \int_0^\infty \left[ e^{-\lambda t} \right] \left[ 1 - e^{-\lambda t} \right] dG(\lambda) \]

seen 0 times in \([-1, 0] \]
not seen 0 times in \([0, t] \)

Notice you can derive expressions for

\[ E[\text{\# seen a times in } [-1, 0] \text{ and b times in } [0, t]] \]

in the straightforward way.
Notice also that you don't need C to evaluate this, just $CdG(\lambda)$.

Q: estimate $\Delta(t)$ given $n_{\infty}$

Notice

$$1 - e^{-\lambda t} = \lambda t - \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!} - \frac{(\lambda t)^4}{4!} \ldots$$

So:

$$\Delta(t) = n_1 t - n_2 t^2 + n_3 t^3 - n_4 t^4 \ldots$$

(assuming convergence, etc).

Natural estimator:

Assume $n_1 = n_1$, $n_2 = n_2$, etc.

Substitute
Paper gives word counts, word type counts

884647 words,
14376 only once
31534 word types

Q: if we saw another 884647 words, how many new words would there be?

A: \[ t = \frac{\# \text{ of words in new}}{\# \text{ in net}} \]
\[ t = 1 \]
\[ \Delta(t) = 11430 = \Delta(1) \]
\[ \text{Var}(\Delta(1)) \text{ by assuming the n}_x \text{ are indep, poisson} \]
\[ \text{Var}\{\Delta(1)\} \approx \sum_{i=1}^{8} n_i t^i = 31534 \]
\[ \text{std} = 178 \]
(skip Fisher model)

Notice that, for \( t > 1 \)

\[
n_{1}t - n_{2}t^{2} + n_{3}t^{3} - \ldots
\]

is a series that oscillates savagely.

You could interpret this several ways:

• it doesn't converge. - panic
• the oscillations "cancel", and we need some way to accelerate this cancellation

It's quite plausible, as \( n \to 0 \)

\[\text{as } x \to \infty.\]

This gives §4 - Euler's transform.

Now skip to §7.
Recall that $\Delta(t)$ may be hard to estimate for large $t$ (because there may be low frequency words).

Instead, they look for a lower bound on $\Delta(t)$ — call this bound $b(t)$.

The problem becomes:

$$b(t) = \inf_{CdG(x)} \left[ \int_0^\infty e^{-\lambda t} [1 - e^{-\lambda t}] [C \cdot dG(x)] \right]$$

Subject to:

$$n_x = \epsilon \int_0^\infty \left[ \frac{x^e}{x!} \right] [CdG(x)]$$

This would be an LP in $CdG(x)$, IF we knew $n_x$. 
Strategy 1:

* assume \( n_x = n_x \)

* discretize \( n_x \)

\[ \implies \text{then we have an honest LP and can solve.} \]

* This works for \( \mathbb{E}[X] + t \), but failed on object data (infeasible - ??)

Strategy 2:

* assume \( \mu + \gamma \sqrt{n_x} < n_x + r \sqrt{n_x} \)

\( n_x \geq n_x - \gamma \sqrt{n_x} \)

(i.e. \( \gamma \) stds away from mean)

* This might slacken the bound, but is probably better practice
But $S_2$ isn't all that reliable either
(on objects, get feasibility only if you apply big $F$ AND use only $n_x$ for $x \in [1..5]$ which is worrying.)

What is going on here?

Consider $\frac{e^{-\lambda}x^c}{c!}$ as a function of $\lambda$

This means we can't have

$\eta_1 = 100; \eta_2 = 0; \eta_3 = 100$; etc.

because these functions overlap so strongly. $\int e^{-\lambda} C_dG(\lambda)$ is similar to $\sum_{i=5}^c e^{-\lambda} \lambda^2 C_dG(\lambda)$
Alternative View:

\[ \text{feasible } C_dG(x) \rightarrow \text{linear map} \] (integals) \[ \rightarrow \eta_x \]

\( \inf \text{ - min; Convex; finite min;} \)

Since \( C_dG(x) \geq 0 \)
this isn't the whole space we \( \eta \) 's can't occur.

This picture strongly implies that use this picture with plots;
there are vectors of \( \eta_x \) that are (a) non-negative (b) infeasible

and if you use \( \eta_x = \eta_x \) this gets infeasibility
Also explains why large $r$ is required, and why large $n_x$ creates problems (the $n_x$ estimates are poor).

What to do?

* we actually know quite a bit about $n_x$, which is what we should be working with
  * approximately Gaussian
  * $\text{Var approx} = \text{mean}$

**Strategy 3**

Assume $r_{n_x} \sim N(n_x, \sqrt{n_x})$

and $\tilde{n}_x = \hat{r}_x$. 
Now we must \( \text{\textcircled{2}} \) estimate the counts and \( \text{\textcircled{3}} \) estimate \( b(t) \)

\[
\inf_{\text{CdG}(\lambda)} \int e^{-\lambda} \left[ 1 - e^{-\lambda t} \right] \text{CdG}(\lambda) + \mu \sum_{i=1}^{K} s_i^2
\]

\[
s_i^2 = \frac{(r_i - n_i)^2}{2n_i^2}
\]

\[
r_i = \int \frac{e^{-\lambda t}}{c(t)} \text{CdG}(\lambda)
\]

Q: how to choose \( \mu \)?
A: cross validation

Now a QP, but it's convex, so no worries.
Notice also that
\[ g(\lambda,t) = e^{-\lambda} (1 - e^{-\lambda t}) \] is important

This (basically) looks for weight in small \( \lambda \)'s (at \( \text{CdG}(\lambda) \)) which makes sense.

**Issue**: lower bounds are helpful, but we want more (estimates, etc).

**Options**
- work w/ continuous \( \text{CdG}(\lambda) \) models
- make models explicitly discrete.
Notice one attractive feature of this formulation.

- \( x \), \( b(t) \), \( \Delta(t) \) are quite insensitive to "small" changes in \( C_0[G(t)] \)

- because

\[
\frac{e^{-x} \cdot x^x}{x!}
\]

This means that big shifts in where the weight is in a prob model are required to change counts.

In turn, we could use quite rough discretizations (but finer near 0).