Then the rvs $(S_i \mid \theta)$ are IID Bernoulli(θ)

and the rv $S = \sum_{i=1}^{n_T} S_i$ has a conditional

binomial dist: $(S \mid \theta) \sim \text{Binomial}(n_T, \theta)$

It's meaningful to talk about the conditional expectation rv. $E(S \mid \theta) = n_T \theta$ (a linear function of θ),

and – via Bayes' Theorem – it's even more meaningful to talk about the conditional expectation <u>rv.</u> $E(\theta \mid S)$ (more about this later)

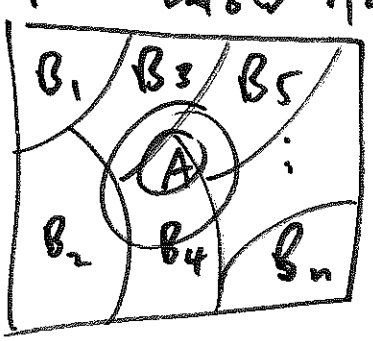
---

and the <u>constant</u> $E(\theta \mid S = s)$.

Important consequence of the def. of conditional expectation

(15 Aug)

Remember the Law of Total Prob.!



$$P(A) = \sum_{i=1}^{n} P(B_i)\, P(A \mid B_i)$$

(LTP)

Continuous version of LTP

X, Y continuous r.v.

for which all named densities exist $\rightarrow$

$$f_Y(y) = \int_{-\infty}^{\infty} f_X(x) \cdot f_{Y|X}(y|x)\, dx$$

Earlier we agreed that, by definition,

$$E(Y|x) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dx$$

So watch the following slightly magical calculation:

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

$$= \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_X(x)\, f_{Y|X}(y|x)\, dx \right] dy$$

ok to interchange order of integration

$$= \int_{-\infty}^{\infty} f_X(x) \left[\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy \right] dx$$

$$= \int_{-\infty}^{\infty} f_X(x) \cdot E(Y \mid x)\, dx \quad (*)$$

, and this is of the form {weighted average of $E(Y|x)$, with $f_X(x)$ as the weights}

Recall that for any continuous r.v. W,

$$E(W) = \int_{-\infty}^{\infty} w\, f_W(w)\, dw$$

and

$$E[h(W)] = \int_{-\infty}^{\infty} h(w)\, f_W(w)\, dw$$

so $(*)$ is just

$$E_X[E(Y \mid X)]$$

and we have shown that (Adam)

$$E(Y) = E_X[E(Y \mid X)]$$

This is referred to as part (1) of the <u>double expectation theorem</u>; strangely, DS don't even mention that name, calling it instead the <u>LTP for expectations</u>.

I need to postpone examples of these conditional expectation calculations until we've covered more standard distributions.

X, Y r.v. such that $f_{Y|X}(y|x)$ exists $\rightarrow$ it makes sense to speak not only of $E(Y|x)$, the <u>mean</u> of $f_{Y|X}(y|x)$, but also of the <u>variance</u> of that dist.

Def $V(Y|x) \triangleq E_X\left\{\left[Y - E(Y|x)\right]^2 | x\right\}$ is called the <u>conditional variance</u> of Y given $X = x$, and the <u>rv</u> $V(Y|X)$ is just $g(X)$, the conditional variance of Y given X.

$= g(x)$

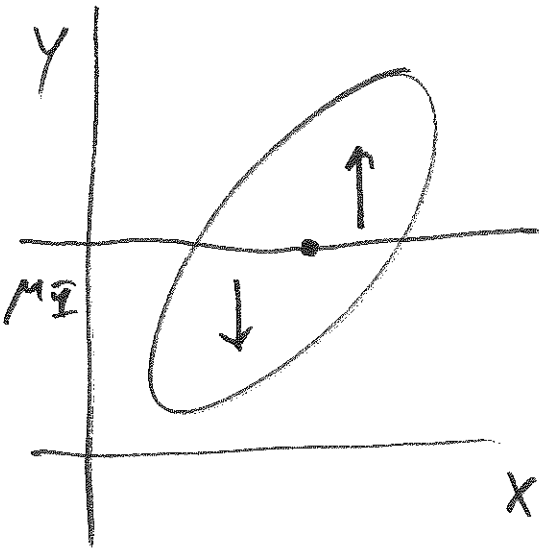
The payoff from all of this (formalizing Galton's intuition)

Theorem X, Y related rv, want to use some function $\hat{Y} = d(X)$ to predict Y from X $\rightarrow$ the prediction $\hat{Y} = d(X)$ that minimizes the MSE $E(Y - \hat{Y})^2 = E\{[Y - d(X)]^2\}$ is $\hat{Y} = d(X) = E(Y|X)$, the conditional expectation of Y given X.

Part 2 of the double expectation theorem

X, Y r.v. such that all of the following expressions exist, $\rightarrow$

$$V(Y) = E_X[V(Y|X)] + V_X[E(Y|X)]. \qquad \text{(Eve)}$$



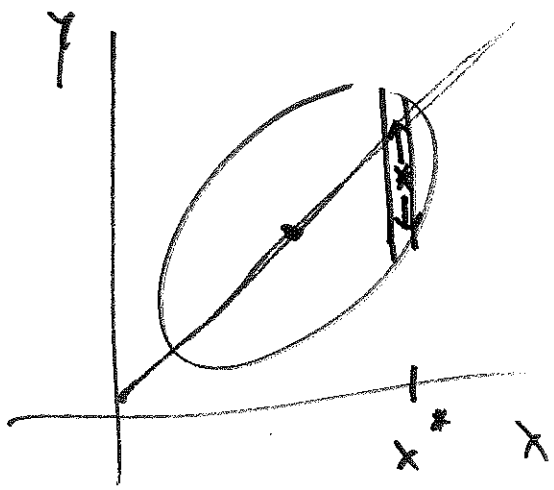
Imagine a 2-part game!

Stage 1 Predict Y without knowing X. Well, if you buy into MSE as your measure of "goodness" of a prediction, we know that you should predict $\hat{Y}_{\text{no } X} = \mu_Y = E(Y)$

and your resulting MSE will be

$$E\left[(Y-\mu_Y)^2\right] = V(Y) = \sigma^2_Y.$$

Stage 2 Observe X, now predict Y



Let's say $X = x^*$. Then we know the MSE-optimal prediction is $\hat{Y}_{X=x^*} = E(Y|X=x^*)$

and your resulting MSE will be

$$E\left\{\left[Y - E(Y \mid X = x^*)\right]^2\right\} = \underbrace{V(Y \mid x^*)}_{(**)}.$$

From the vantage point of someone thinking about stage 2 before it happens, X is not yet known, so the expected value of $(**)$, namely $E_X[V(Y \mid X)]$, is the best you can do to guess at how good the stage 2 prediction will be.

The second part of the double expectation theorem says

$$\underbrace{V(Y)}_{\substack{\text{MSE of} \\ \hat{Y}_{no\,X}}} = \underbrace{E_X[V(Y \mid X)]}_{\substack{\text{"E(MSE)" of} \\ \hat{Y}_X = E(Y \mid X)}} + V_X[E(Y \mid X)]$$

But since variances are always non-negative,

$$V_X[E(Y|X)] \geq 0 \text{ , so}$$

$$\underbrace{E_X[V(Y|X)] + V_X[E(Y|X)]}_{V(Y)} \geq \underbrace{E_X[V(Y|X)]}_{\text{"E(MSE) of } \hat{Y}_X\text{"}}$$

$$V(Y) \geq \text{MSE of } \hat{Y}_{\text{no } X}$$

Thus you always expect your predictive accuracy to get better (or at least stay the same) when you use $E(Y|X)$ to predict Y.

Another complete switch in subject

Utility

Q: How to take action sensibly when the consequences are uncertain?

A: There is a theory of optimal action under uncertainty; it's called <u>Bayesian decision theory</u> — a concept called <u>utility</u> is central to this theory. The theory takes its simplest form when comparing gambles

Example $\underline{X}$ has discrete PF $f_X(x) = \begin{cases} \frac{1}{2} & x = -\$350 \\ \frac{1}{2} & x = +\$500 \\ 0 & \text{else} \end{cases}$

Suppose $\underline{X}$ = your net gain from gamble (A),

$\underline{Y}$ has discrete PF $f_Y(y) = \begin{cases} \frac{1}{3} & y = 40 \\ \frac{1}{3} & y = 50 \\ \frac{1}{3} & y = 60 \\ 0 & \text{else} \end{cases}$

and $\underline{Y}$ = your net gain from gamble (B).

Turns out that $E(\underline{X}) = \$+75$, $E(\underline{Y}) = \$+50$ so is (A) automatically better than (B)?

Note that with (B) you're guaranteed to win at least 840, while (A) has no such guarantee; is (A) still automatically better $\boxed{\text{for you}}$ than (B)? A risk-averse person would grab (B) quickly; a risk-seeking person would probably pick (A).

Evidently something more than just computing $E(X)$, $E(Z)$ is going on.

DS def. of utility function	<u>Your</u> utility function $u(x)$ is that function which assigns to each possible net gain $-\infty < x < \infty$ a real # $u(x)$ representing the <u>value</u> <u>to you</u> of gaining x.

Q: If x is money, why not just use $u(x) = x$? (utility linear in money)

A: Lovely, subtle answer first supplied by Daniel Bernoulli (1700 – 1782), (Swiss mathematician) related to Jacob Bernoulli (1654 – 1705), for whom the Bernoulli distribution was named.

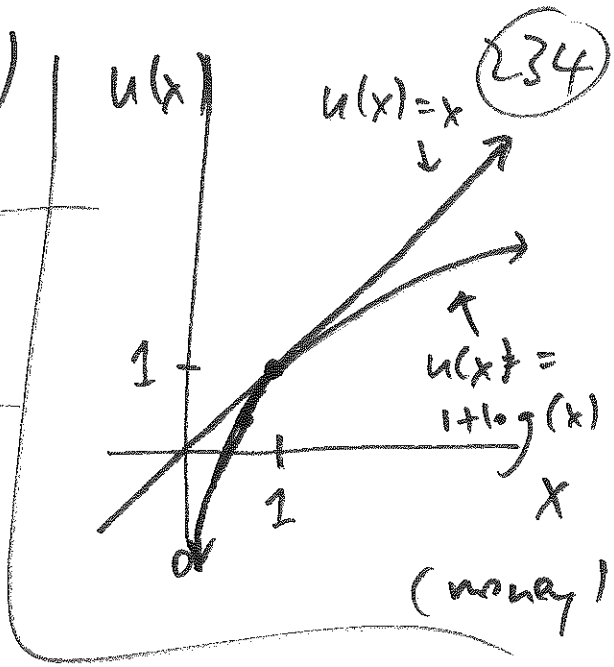
Daniel B: If your entire net worth is (say) \$10, then the <u>value</u> to you of a new \$1 is much greater than if your entire net worth is (say) \$1,000,000; thus the utility of money is <u>sublinear</u> (meaning that it doesn't grow with x as fast as $f(x) = x$ does)

Daniel B proposed one particular sublinear function for <u>utility</u>,

namely $u(x) = 1 + \log(x)$ (for $x > 0$)

(Daniel B also invented the word <u>utility</u>)

(Although the idea goes back at least to Aristotle (384-322 BCE))



<u>Definition</u>

(Principle of Expected Utility Maximization)

You are said to choose between gambles by <u>maximizing</u> <u>expected utility</u> (MEU) if, with $u(x)$ your utility function,

(a) you prefer gamble X to gamble Y if $E[u(X)] > E[u(Y)]$ and (b) you're indifferent between X and Y if $E[u(X)] = E[u(Y)]$

MEU first explored in depth by British

{mathematician, philopher, economist} Frank Ramsey (1903 - 1930), who died at 26 of liver failure.

Theorem (von Neumann - Morgenstern (1947))

Under 4 reasonable axioms, MEU is the best you can do.

John von Neumann (1903 - 1957)

Hungarian - American

{mathematician, physicist, computer scientist, ⋮}

died at 53 of cancer

Simple example You bought a single $2 ticket in the Powerball lottery examined in Homework 1 problem 2: the drawing on 30 Jul 2016 for which the Grand Prize was $487 million. Let X be the unknown amount you will win

Oskar Morgenstern (1902 - 1977)

German-economist American

(thinking about X before the drawing).

$\boxed{Q:}$ Before the drawing, someone offers you $\$x_0$ for your ticket; should you sell?

$\boxed{A:}$ With $u(x)$ as your utility function, your expected gain if you keep the ticket is $E[u(X)]$; if for you $u(x) = x$ (utility $\triangleq$ money) then $E(u(X)) = \$1.99$

Action 1 (sell): you gain $\$x_0$ for sure

Action 2 (keep): your expected utility is $E[u(X)]$

Under MEU you should sell if $\boxed{u(x_0) > E[u(X)]}$

If $u(x) = x$ for you then your optimal action is (sell if offered more than \$1.99).

236

Match	x	$P(X=x)$	$x \cdot P(X=x)$
5w, 1R	\$487,000,000	$\frac{1}{292,201,338}$	\$1.667
5w, 0R	\$1,000,000	$\frac{1}{11,688,053.52}$	0.086
4w, 1R	\$50,000	$\frac{1}{913,129.18}$	0.055
4w, 0R	\$100	$1/36,525.17$	0.003
3w, 1R	\$100	$1/14,494.11$	0.007
3w, 0R	\$7	$1/579.76$	0.012
2w, 1R	\$7	$\frac{1}{701.33}$	0.010
1w, 1R	\$4	$\frac{1}{91.98}$	0.043
0w, 1R	\$4	$\frac{1}{38.32}$	0.104
			$\boxed{\$1.99 \ (!)}$

X has 9 possible values x (discrete),

so $E(X) = \sum_{\text{all 9 possibilities}} x \cdot P(X=x) = \$1.99$.

Related but different problem

On the 13 Jan 2016 drawing the Powerball jackpot was $1.6 billion. (238)

X = your winnings

X uncertain before the drawing

Redo calculation on p. (236): $E(X)$ is now $5.80 on a $2 ticket

new 1st row in table is $\frac{1{,}600{,}000{,}000}{292{,}201{,}338} \doteq 85.476$

Q: If $u(x) = x$ for you, is it rational to sell all your assets & buy as many lottery tickets as possible? (*)

A: Yes, but that's a silly utility function; to be realistic you'd have to subtract from x the

monetary value to you of the disruption of your life that would ensue with action

(*) A catalog of useful distributions

(DSch.5) Case 1: Discrete — Bernoulli

$X \sim \text{Bernoulli}(p)$, $0 < p < 1$, if

$$f_X(x) = p^x (1-p)^{1-x} \overbrace{I_{\{0,1\}}(x)}^{\text{support}(X)}$$

$$= \begin{cases} p & \text{for } x = 1 \\ 1-p & \quad\quad 0 \\ 0 & \text{else} \end{cases}$$

$E(X) = p$

$V(X) = p(1-p)$

$\psi_X(t) = pe^t + (1-p)$ for all $-\infty < t < \infty$

$SD(X) = \sqrt{p(1-p)}$

Def If the X_i in $X_1, X_2, \ldots$ are IID Bernoulli (p), then $(X_1, X_2, \ldots)$ are called <u>Bernoulli trials</u> with <u>parameter</u> p; if the sequence $(X_1, X_2, \ldots)$ is infinite this defines a <u>Bernoulli (stochastic) process</u>.

Binomial $X \sim \text{Binomial}(n, p)$ (i.e., X follows the Binomial distribution with parameters n (positive integer) and $0 < p < 1$)

$$\leftrightarrow f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \overbrace{I_{\{0,1,\ldots,n\}}(x)}^{\text{Support}(X)}$$

Consequences ① $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$

$$\rightarrow X = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$$

$X \sim \text{Binomial}(n, p)$ | $E(X) = n \cdot p$ | $V(X) = np(1-p)$

$\psi_X(t) = [pe^t + (1-p)]^n$ for all $-\infty < t < \infty$

$SD(X) = \sqrt{np(1-p)}$

Example | Cortaneda v. Partida (1977)

Grand juries in the U.S. judicial system have <u>catchment areas</u>: everybody 18 & over living in the judicial district for that grand jury (& a few other minor restrictions)

Hidalgo county, Texas

eligible pool was 79.1% Mexican-American

$2\frac{1}{2}$ yr period at issue in Supreme Court case: 220 people called to serve on grand juries, but only 100 of them are Mexican-American

extreme southern border of TX with Mexico

$\boxed{Q:}$ <u>Prima facie</u> case of discrimination?

Before this $2\frac{1}{2}$ yr period, let X be your prediction of # of Mexican-Americans among the 220 people

If no discrimination,

$X \sim \text{Binomial}(220, 0.791)$ $(X \mid T_1)$

T_1 = theory 1 = no discrimination

$E(X \mid T_1) = \underbrace{(220)}_{n}\underbrace{(0.791)}_{p} \doteq 174.0$

$SD(X \mid T_1) = \sqrt{np(1-p)} \doteq 6.0$

Q: If you were expecting 174 give or take 6, would you be surprised to see 100?

A: You'd be astonished

Frequentist statistical answer	$P(X \le 100 \mid T_1) = 8.0 \cdot 10^{-28}$ so T_1 looks ridiculous
Bayesian statistical answer	Need to compute $P(T_1 \mid X = 100)$, not the other way around (later)

Hypergeometric | A finite population has A elements of type 1 and B elements of type 2; total population size (A+B).

You choose n elements at random without replacement from this population (i.e., you take a simple random sample (SRS) of size n)

Let X = (# elements of type 1 in your sample)

Then (as noted in homework 1 problem 2) X follows the hypergeometric distribution with parameters (A, B, n).

As we saw in that problem, the PF of X is

$$f_X(x \mid A, B, n) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} \underbrace{I[\max\{0, n-B\} \le x \le \min\{n, A\}]}_{\text{support}(X)}$$

for (A, B, n) non-negative integers with $n \le A+B$

Consequences

① $E(X) = n \cdot \frac{A}{A+B}$

② $V(X) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{A+B-n}{A+B-1}\right)$

Note that if your sampling had been <u>with</u> replacement (i.e., you take an IID sample), X would have been Binomial with the same value of n and $p = \frac{A}{A+B}$; in that case $E(X) = np = n\frac{A}{A+B}$ and

$V(X) = np(1-p) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$. (compare)

If you let $T = (A+B)$ be the total # of elements in the population,

sampling method	mean	variance
with repl. (IID)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$
without repl. (SRS)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{T-n}{T-1}\right)$

$0 \le \alpha = \frac{T-n}{T-1} \le 1$ is called the <u>finite population correction</u>

3 special cases worth considering

(a) $(n=1)$ $\alpha = 1 \leftrightarrow$ SRS = IID with only 1 element sampled

(b) $(n=T)$ $\alpha = 0 \leftrightarrow$ If you exhaust the entire population in SRS, you have no uncertainty

(c) (n fixed, $T \uparrow$) $\alpha \uparrow 1$ [redacted] $\leftrightarrow$ with a small sample from a large population,

SRS $\doteq$ IID

Poisson ($\lambda > 0$) $X \sim \text{Poisson}(\lambda)$

$\leftrightarrow$ X has PF $f_X(x) = \dfrac{\lambda^x e^{-\lambda}}{x!} I_{\{0,1,\ldots\}}(x)$

$E(X) = \lambda$

$V(X) = \lambda$

Thus for the Poisson dist.

$\dfrac{V(X)}{E(X)} = 1$

Def. If $E(X)$ and $V(X)$ both exist and $E(X) \neq 0$, $\dfrac{V(X)}{E(X)}$ is called the variance-to-mean ratio (VTMR)

$\psi_X(t) = e^{\lambda(e^t - 1)}$

$-\infty < t < \infty$

The Poisson can be unrealistic as a consequence of its VTMR of 1, because

many rvs that represent counts of occurrences of events in time intervals of fixed length have VTMR > 1.

The Poisson & Binomial distributions both count the numbers of "successes" in a process unfolding in time, so it should not be surprising to find out that these 2 dist. are related:

---

when $\begin{pmatrix} n \text{ is large} \\ p \text{ is close to } 0 \end{pmatrix}$, $\text{Binomial}(n,p) \doteq \text{Poisson}(n \cdot p)$

<u>Theorem</u> n positive integer, $0 < p < 1$ $X \sim \text{Binomial}(n,p)$

$\lambda > 0$, $Y \sim \text{Poisson}(\lambda)$ / Choose any sequence

$\{p_n\}_{n=1}^{\infty}$ of values between 0 and 1 with

$\lim_{n\to\infty} n \cdot p_n = \lambda$ then $f_X(x \mid n, p_n) \xrightarrow[n\to\infty]{} f_Y(y \mid \lambda)$

Poisson process, revisited

Def

A poisson process (or space, or volume, or...) with rate λ per unit time is a stochastic process with two properties:

(a) # arrivals in every interval of time of length $t \sim \text{Poisson}(\lambda t)$

(b) #s of arrivals in all disjoint (non-overlapping) time intervals are independent

Example

~~Parasitic~~ protozoa in drinking water

There's a kind of parasitic

organism called cryptosporidium that's capable of getting into the public drinking water supplies; at one stage in their life cycle they're called oocysts. They can make people sick at a concentration of only 1 oocyst per 5 liters ≐ 1.3 gallons of water. One problem is that it can be hard to detect these oocysts with water filtration. Suppose that, in the water supply of your city, oocysts occur according to a Poisson process with rate λ oocysts per liter, & that the filtering system your water utility company uses can capture all the oocysts in a water sample but only has

probability p of detecting each oocyst that's actually there. (& counting events are independent)

Let Y = actual # oocysts in t liters of water, and $X_i = \begin{cases} 1 & \text{if oocyst } i \text{ gets counted} \\ 0 & \text{else} \end{cases}$

X = # counted oocysts | then $(X \mid Y = y) = \sum_{i=1}^{y} X_i$

under these assumptions, $(X \mid Y = y) \sim \text{Binomial}(y, p)$

Q: What's the dist. of X? | A | By the law of total probability

$$f_X(x) = P(X = x) = \sum_{y=0}^{\infty} P(Y = y) P(X = x \mid Y = y)$$

for all $x = 0, 1, \ldots$

in which $P(Y = y) = \dfrac{(\lambda t)^y e^{-\lambda t}}{y!}$ for $y = 0, 1, \ldots$

and $P(X=x \mid Y=y) = \binom{y}{x} p^x (1-p)^{y-x}$

Notice that if $X=x$, $Y \geq x$ because the actual number of oocysts (Y) has to be at least as large as the number of oocysts detected (X).

After a careful calculation you get:

$$f_X(x) = \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \frac{(\lambda t)^y e^{-\lambda t}}{y!}$$

$$= \frac{e^{-p\lambda t} (p\lambda t)^x}{x!}$$

i.e., $X \sim \text{Poisson}(p\lambda t)$: losing a proportion $(1-p)$ of the oocysts to faulty counting just lowers the rate of the Poisson process from λ/liter to $\lambda \cdot p$/liter (makes excellent sense).

In practice oocysts are hard to detect: p is small (not far from 0). $\boxed{Q:}$ How much water (t liters) do you need to filter to achieve $P(\text{at least 1 oocyte detected}) \geq 1-\alpha$ for small α? $\boxed{A:}$ Not hard to work out

$$P(\text{at least 1 detected}) = 1 - P(\text{none detected})$$

$$= 1 - P(X=0) = 1 - e^{-p\lambda t} \geq 1-\alpha$$

$$\Leftrightarrow \alpha \geq e^{-p\lambda t} \Leftrightarrow \ln\alpha \geq -p\lambda t \Leftrightarrow$$

$$\boxed{t \geq -\frac{\ln\alpha}{p\lambda}}$$

Example $\alpha = .01$, $p = 0.1$, $\lambda = 0.2$ / liter (1 per 5 liters) $\updownarrow$ minimum sickness level

to achieve prob 99%, t has to be at least 230.3 liters. (17 Aug 2018)