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).

Negative Binomial Distribution

You're watching a potentially endless sequence of Bernoulli trials with constant success probability p.

Let X = # failures before r^{th} success

r integer ≥ 1

You can show that X follows what's called the Negative Binomial dist:

its PMF is $f(x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x$.

with parameters (r, p) $(0 < p < 1)$

X

The name comes $I_{\{0,1,2,\ldots\}}(x)$.

from the fact that, when you watch a sequence of Bernoulli trials with unknown constant success probability p unfold, there are two different ways to

estimate p: decide ahead of time to sample n (known constant) success/failure trials, and record the (random) # S of successes you see (from which a reasonable estimate would be $\hat{p}_B = \frac{S}{n}$ ← Binomial).

(or) decide ahead of time that you're going to sample until you've seen s (known constant) successes & record the (random) # of trials N needed to accumulate that many successes (from which a reasonable estimate would be $\hat{p}_{NB} = \frac{s}{N}$ ← Negative Binomial).

Special Case of Negative Binomial

Set $r=1$ and record the number $\underline{X}$ of failures until the first success: $\underline{X}$ is said to follow the

Geometric (p) distribution, with

$$f_{\underline{X}}(x \mid p) = p(1-p)^x \, \underline{1}_{\{0,1,\ldots\}}(x)$$

(parameter p)

support of $\underline{X}$

Consequence $\underline{X}_1, \ldots, \underline{X}_n$ IID Geometric(p)

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

This is a direct analogue to the Bernoulli/Binomial story: $\underline{X}_1, \ldots, \underline{X}_n$ IID Bernoulli(p) $\rightarrow \sum_{i=1}^{n} \underline{X}_i \sim \text{Binomial}(n, p)$.

$X \sim$ Negative Binomial (r, p)

$$\psi_X(t) = \left[\frac{p}{1-(1-p)e^t}\right]^r \quad \text{for } t < \log\left(\frac{1}{1-p}\right)$$

from which $E(X) = \dfrac{r(1-p)}{p}$, $V(X) = \dfrac{r(1-p)}{p^2}$

Consequence $X \sim$ Geometric (p) $\rightarrow$

$\left\{\begin{matrix} k \\ t \end{matrix}\right\}$ both non-negative integers

$$P(X = k+t \mid X \geq k) = P(X = t)$$

this is called the memoryless property of the Geometric distribution, and it turns out that this is the only

discrete distribution with this property.

F F F F F S

X = # failures until first success = 5 (here)

Y = # failures, starting at trial $(k+1 = 2$ here$)$ until next success $(= 4$ here$)$

Then Y has the same dist. as X and is independent of what happened on the first k trials, i.e., "the process has no memory".

Important Core 2: Continuous Distributions

Normal (Gaussian) Distribution

$X \sim \text{Normal}(\mu, \sigma^2)$ mean μ, variance $0 < \sigma^2 < \infty$

PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

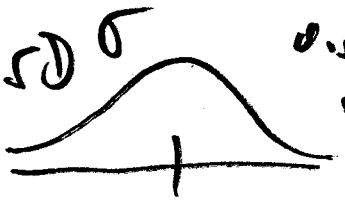
The Normal dist. is the single most important dist. in all of probability & statistics, mainly for 2 reasons: ① many observable random processes have dist. shapes that are close to the "bell curve" (Normal PDF), and ② the Central Limit Theorem (CLT), which we'll examine soon.

Properties of the Normal Dist.

$X \sim \text{Normal}(\mu, \sigma^2)$ $\quad (N(\mu, \sigma^2))$ $\quad E(X) = \mu$

$V(X) = \sigma^2$, $SD(X) = \sigma$

$$\psi_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$



dist. of X / (center of symmetry: mean, median, mode) all $= \mu$

Consequences ① $X \sim \text{Normal}(\mu, \sigma^2)$,

$Y = aX + b$, $\left(\begin{matrix} a \neq 0 \\ b \end{matrix}\right)$ fixed constants $\rightarrow$

$$Y \sim \text{Normal}(a\mu + b, a^2\sigma^2).$$

In other words, Normality is preserved under linear transformation

Def.

The Normal dist. with mean $\mu = 0$ and SD $\sigma = 1$ is called the standard Normal dist.

The PDF of $Z \sim \text{Normal}(0, 1)$ is

$$\phi_Z(x) \triangleq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$ and its

ϕ: phi (lower-case)

CDF is $\Phi(x) \triangleq \int_{-\infty}^{x} \phi_Z(t)\, dt$

Φ: upper-case phi; Z

It turns out that e^{-cx^2} $(c>0)$ has no anti-derivative in closed form, so $\Phi(x)$ cannot be summarized in a formula; instead it's approximated by numerical integration (see p. 861 in JS).

Consequences, continued

② Because the Normal PDF (for all $x \in \mathbb{R}$) is symmetric, $\Phi(-x) = 1 - \Phi(x)$

and $\Phi^{-1}(p) = -\Phi^{-1}(1-p)$ (for all $0<p<1$)

---

③ $X \sim \text{Normal}(\mu, \sigma^2) \rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$,

so that $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

and $F_X^{-1}(p) = \mu + \sigma \Phi^{-1}(p)$

Empirical Rule

Part 1 start at the mean (μ) of a distribution and go 1 SD either way: you will find (about $\frac{2}{3}$) (68%) of the probability in the interval ($\mu \pm 1\sigma$)

Part 2 Ditto 2 SDs either way: ($\mu \pm 2\sigma$) capturing (about most 95%) of the probability

Part 3 Ditto 3 SDs either way: ($\mu \pm 3\sigma$) capturing almost all (99.7%) of the probability

This Rule is <u>exact</u> for all Normal dists & is a surprisingly

good approximation for many other distributions. This permits an easy trick that's helpful in computing Normal probabilities.

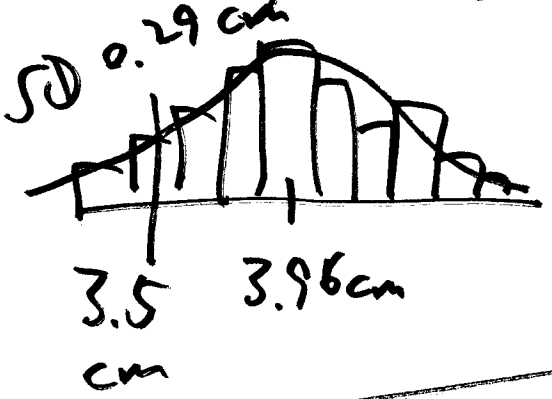
Example: You have a random sample of $n = 103$ immature monarch butterflies, and you measure their wing lengths:

y = wing length (cm)

$$\begin{bmatrix} y_1 = 4.1 \\ y_2 = 3.3 \\ \vdots \\ y_n = 4.7 \end{bmatrix} \quad n = 103$$

mean $\bar{y} = 3.96$ cm

SD $s = 0.29$ cm



histogram of wing length

Q: About what % of the sampled butterflies had wing length ≤ 3.5 cm?

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

sample mean — sample SD

sorted y

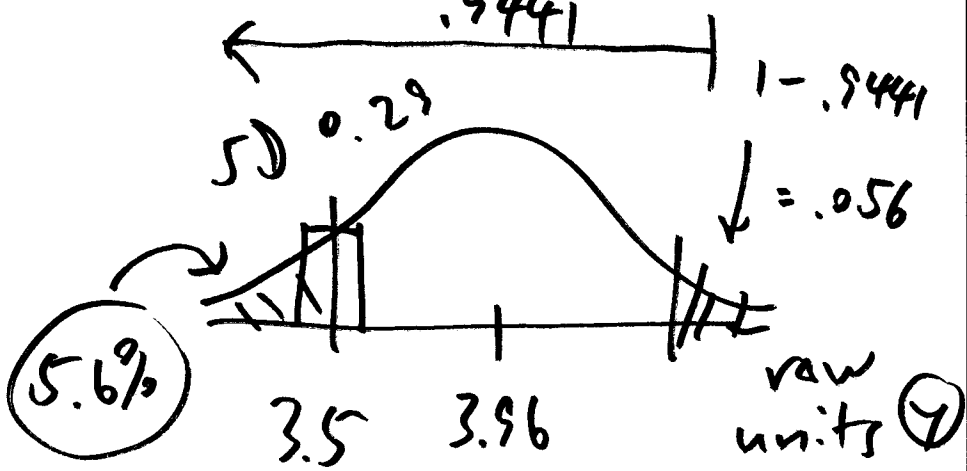
3.2
3.3
⋮
3.5
3.5
3.5
3.5
3.6
⋮
4.7

8

103

A_1 (exact) $\frac{8}{103} \doteq 7.8\%$

A_2 (approximate)



$$\frac{3.5 - 3.96}{0.29}$$

converting to z_4 for data:

$$z = \frac{y - \bar{y}}{s} = z_4$$

for random variables

$$Z_4 = \frac{Y - \mu}{\sigma} = z_4$$

keeping track of histogram bar edges: continuity correction

More consequences

(4) $X_1, \dots, X_k$ independent,

$X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$

$\Rightarrow \sum_{i=1}^{k} X_i \sim \text{Normal}\left(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2\right)$

~~nice additive property~~

this is why Normal dists are indexed by variance rather than SD.

Notation: $\text{Normal}(\mu, \sigma^2) \triangleq N(\mu, \sigma^2)$

Example: Population of adult U.S. women: height follows $N(\mu = 65.0\text{in}, \sigma^2 = 3.2^2\text{in}^2)$ dist. ($\sigma = 3.2$ in)

Pop. of adult U.S. men: height follows $N(\mu = 69.5\text{in}, \sigma^2 = 3.3^2\text{in}^2)$ dist.

1 woman chosen at random, height $\underline{W}$;
1 man chosen at random (independently), height $\underline{M}$; P(woman taller than man)

$= P(\underline{W} > \underline{M})$

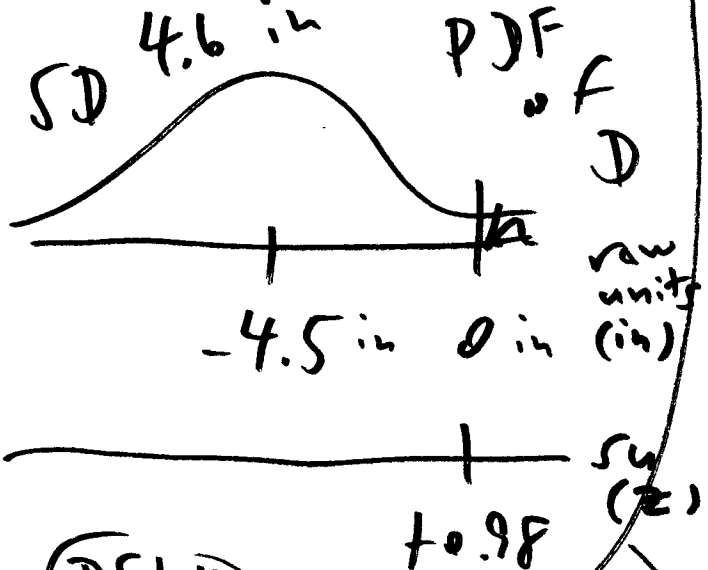
$= ?$

Define $D = W - M$

By consequence ④, $D \sim N($ $65 - 69.5 = -4.5$ in, $3.2^2 + 3.3^2 = 21.1$ in^2 $)$

$P(\underline{W} > \underline{M}) = P(D > 0)$

SD $\sqrt{21.1 \text{ in}^2}$ $= 4.6$ in



Convert to SU:

$$\frac{0_{in} - (-4.5_{in})}{4.6_{in}} = +0.98$$

DS table: .8365

.1635

(25 May 19)

So $P(\underline{W} > \underline{M}) \approx 16\%$

(about 1 in 6)

Def rv $X_1, \ldots, X_n$ → sample mean of $(X_1, \ldots, X_n)$ is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Consequences, continued

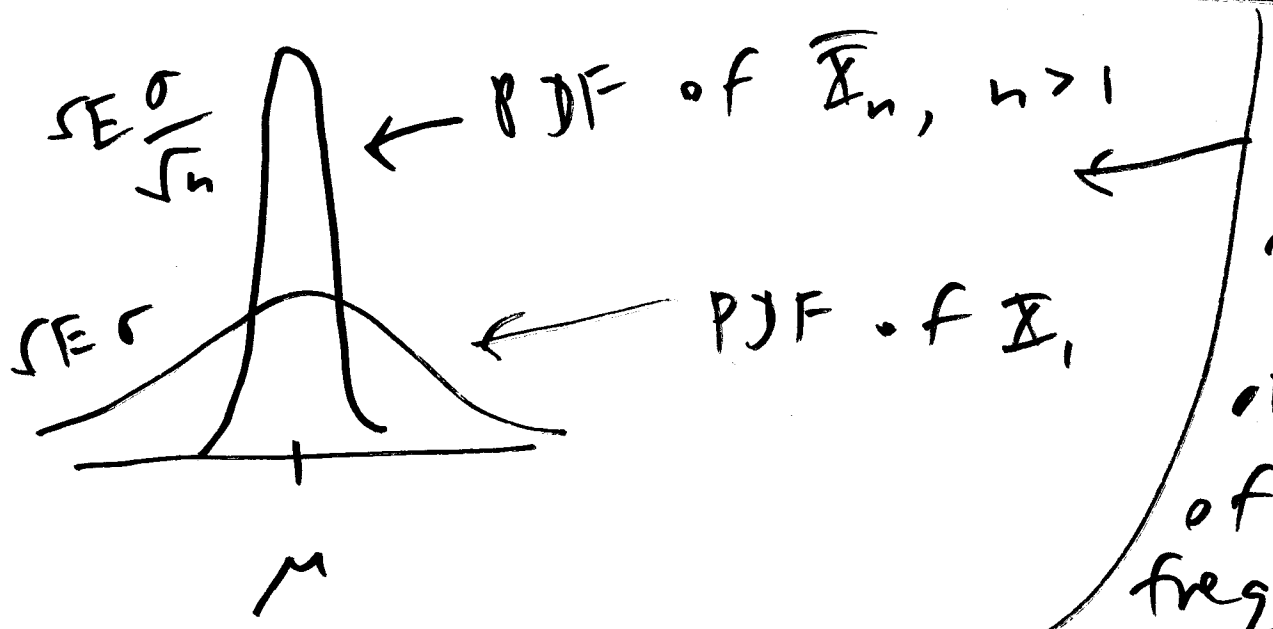
(5) $\left\{ X_i \overset{IID}{\sim} N(\mu, \sigma^2) \quad (i = 1, \ldots, n) \right\}$

→ $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ so $SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$

Because $E(\bar{X}_n) = \mu$, $\bar{X}_n$ is an unbiased estimator of μ

Def In frequentist statistics, the standard deviation (SD) of an estimator $\hat{\theta}$ (rv) of a parameter θ is called the standard error $SE(\hat{\theta})$ of $\hat{\theta}$.

So if you use $\bar{X}_n$ as an estimate of μ, $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \to 0$ as $n \to \infty$



This is the basis of most of frequentist statistical inference

As $n \uparrow$, $\bar{X}_n$ gets better as an estimate of μ, at a $\sqrt{n}$ rate; this is called the Square Root Law.

Unfortunately, this means that to cut the $SE(\bar{X}_n)$ in half, (23 Aug 19) you have to quadruple the sample size.

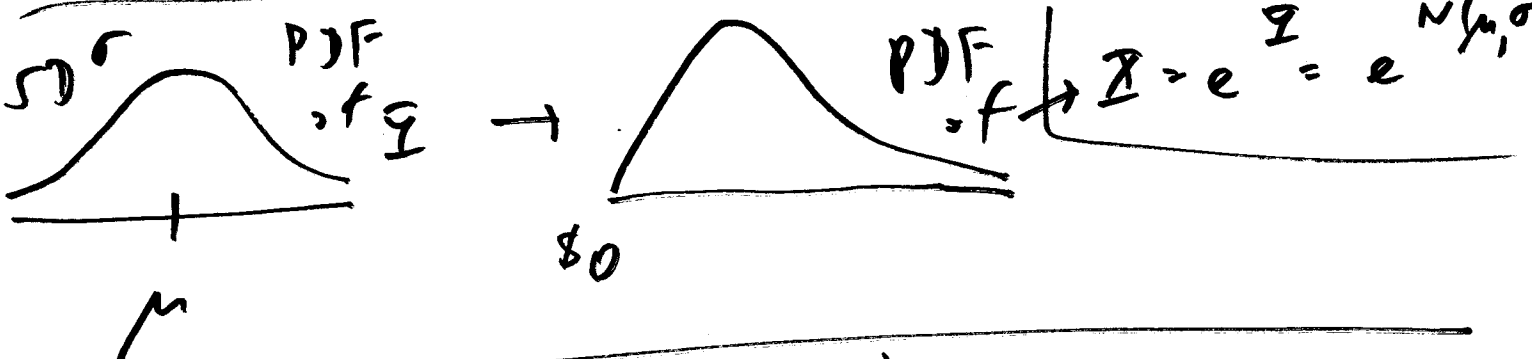
Log Normal Distribution

(This distribution is mis-named ('1900's); it should be called the Exponential-Normal distribution, but we're stuck with a bad name.)

Def.

$X > 0$

If $Y = \log(X) \sim N(\mu, \sigma^2)$, people say that $X \sim \underline{\text{Log Normal}}(\mu, \sigma^2)$.



$X \sim \text{Log Normal}(\mu, \sigma^2)$

- $Y = \log(X) \sim N(\mu, \sigma^2)$

Can get MGF of X from MGF of Y

MGF of X is $\psi_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$

But by definition $\psi_X(t) = E\left(e^{tX}\right) = E\left(e^{t \log X}\right)$

$= E(X^t)$, so we can read the moments of X directly from the MGF of X

$E(X) = \psi_X(1)$

$= \exp\left(\mu + \frac{\sigma^2}{2}\right)$

$V(X) = \psi_X(2) - \left[\psi_X(1)\right]^2$

$= \exp(2\mu + \sigma^2)\left[e^{\sigma^2} - 1\right]$.

Famous case study

Pricing stock options, continued

I shape of a stock, current (known constant) price S_0. Heroic assumption: price

u time units in the future will be

$$S_u = S_0 e^{Z_u}, \quad Z_u \sim N(\mu u, \sigma^2 u).$$

Can write $S_0 e^{Z_u} = e^{Z_u + \log(S_0)}$. Now

$$\left[Z_u + \log(S_0)\right] \sim N(\mu u + \log(S_0), \sigma^2 u),$$

so $S_u \sim \text{LogNormal}[\mu u + \log(S_0), \sigma^2 u]$.

Consider a single time horizon u;

heroic assumption reworded $\rightarrow$ $S_u = S_0 \exp\left[\mu u + (\sigma\sqrt{u}) \cdot Z_1\right]$, $Z_1 \sim N(0,1)$

we need to price the option to buy 1 share of this stock for price q at time u.

Use risk-neutral pricing or in the previous discussion: force present value

$E(S_u) \triangleq S_0$.

Let time scale of u be in years; let the risk-free (continuous-compounding) interest rate be r/year;

then present value of S_u is $e^{-ru} \cdot E(S_u)$.

But by heroic lognormal assumption,

$E(S_u) = S_0 \exp\left(\mu u + \frac{\sigma^2 u}{2}\right)$

So set S_0 equal to

$e^{-ru} S_0 \exp\left(\mu u + \frac{\sigma^2 u}{2}\right)$

result is $\mu = r - \frac{\sigma^2}{2}$

for risk-neutral pricing.

Value of option at time u will be $h(S_u)$, where $h(s) = \begin{cases} s-\xi & \text{if } s > \xi \\ 0 & \text{else} \end{cases}$.

with $\mu = r - \frac{\sigma^2}{2}$, $h(S_u) > 0$ iff

$$Z_1 > \frac{\log\left(\frac{\xi}{S_0}\right) - \left(r - \frac{\sigma^2}{2}\right)u}{\sigma\sqrt{u}} \triangleq c$$

Now a nasty integral

arises: risk-neutral price of option is the present value of $E\left[h(S_u)\right]$,

which is $e^{-ru} E\left[h(S_u)\right] = e^{-ru} \int_c^\infty \left[S_0 e^{\left(r-\frac{\sigma^2}{2}\right)u + \sigma z\sqrt{u}} - \xi\right] \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$.

Careful calculation reveals the following (famous) formula:

$$\boxed{S_0 \Phi(\sigma\sqrt{u} - c) - q e^{-ru} \Phi(-c)}$$ is

the risk-neutral price of the option,

where $c = \dfrac{\log\left(\frac{q}{S_0}\right) - \left(r - \frac{\sigma^2}{2}\right)u}{\sigma\sqrt{u}}$ ← This formula was derived in 1973 by

(Black-Scholes formula)

Gamma Distribution

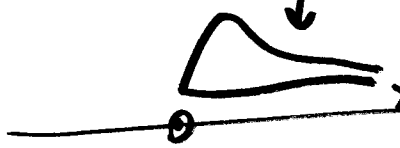
($\alpha, \beta > 0$) X has the Gamma dist. with parameters (α, β), with $X \sim \Gamma(\alpha, \beta)$ or $X \sim \text{Gamma}(\alpha, \beta)$ → X continuous on $(0, \infty)$ with

(American economist) → Fischer Black (1938-1995) died and

Canadian-American economist → Myron Scholes (1941-) ↑ won Nobel Prize in Economics for this work in 1997, together with Robert Merton (1944-) 2003

(age 57 throat cancer)

PDF $f_X(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \underbrace{I(x>0)}_{\text{support of } X}$ ⊛



α is called a shape parameter in the $\Gamma(\alpha,\beta)$ family because it governs things like skewness of the dist.

β is related to the scale of the distribution, which measures how spread out the dist. is

$\Gamma(\alpha)$ is the Gamma function, invented to deal with integrals of functions like ⊛ above:

$$\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1} e^{-x} dx$$

has no anti-derivative in closed form

$\Gamma(\alpha)$ turns out to be a continuous generalization of the factorial function, because $\left(\begin{matrix} n \text{ positive} \\ \text{integer} \end{matrix}\right) \rightarrow \Gamma(n) = (n-1)!$

$\Gamma(\alpha) \rightarrow \infty$ really quickly as $\alpha \rightarrow \infty$, so it's better to evaluate the Gamma PDF on the log scale and then exponentiate:

$$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \exp\left[\alpha \ln(\beta) - \ln\Gamma(\alpha) + (\alpha-1)\ln(x) - \beta x \right]$$

Another way to tame $\Gamma(x)$ is with a Stirling's approximation: $\Gamma(x) \doteq \sqrt{2\pi}\, x^{x-\frac{1}{2}} e^{-x}$ for large x

so that $\ln \Gamma(x) \doteq \frac{1}{2}\ln(2\pi) + (x-\frac{1}{2})\ln x - x$.

$X \sim \Gamma(\alpha, \beta)$ | $\psi_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$ for $t < \beta$

so $E(X) = \frac{\alpha}{\beta}$ | and $V(X) = \frac{\alpha}{\beta^2}$ | $SD(X) = \frac{\sqrt{\alpha}}{\beta}$

Alternative expression: $\psi_X(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$ for $t < \beta$

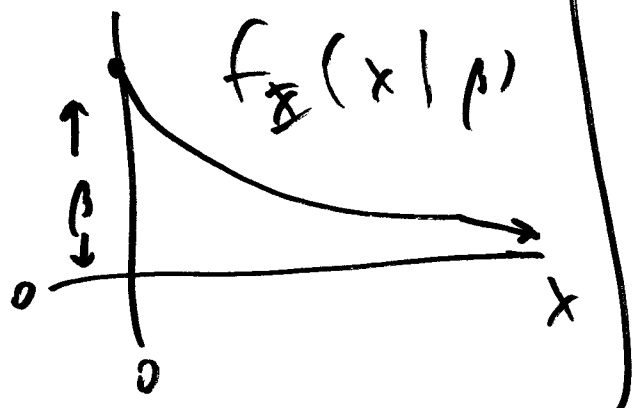
Special case of $\Gamma(\alpha, \beta)$: with $\alpha = 1$ the PDF is

$$f_X(x|\beta) = \beta e^{-\beta x} I(x > 0)$$

But this is just our old friend the Exponential distribution.

$X \sim \text{Exponential}(\beta)$ $\quad \psi_X(t) = \frac{\beta}{\beta - t}, \; t < \beta$

$E(X) = \frac{1}{\beta}$ $\quad V(X) = \frac{1}{\beta^2}$ $\quad SD(X) = \frac{1}{\beta}$



Notice that the Exponential distribution has a TMR equal to β; this suggests it's related somehow to the Poisson dist.

Theorem Suppose that arrivals (events) occur according to a Poisson process with rate β per unit time. Set T_k = time until k^{th} arrival, $k = 1, 2, \ldots$

and define $X_1 = T_1 - 0$

$X_2 = T_2 - T_1$

$\ldots$ $X_k = T_k - T_{k-1}$ for $k = 2, 3, \ldots$

The T_i are called the inter-arrival times.

Then it turns out that $T_i \overset{IFD}{\sim}$ Exponential(β)

The Exponential dist. is also related to the Geometric dist., in that they both have a memory less property

Theorem

$X \sim$ Exponential(β); $t > 0$, $h > 0$

$\rightarrow P(X \geq t+h \mid X \geq t) = P(X \geq h)$

Example X = time (from initial use) until a manufactured product fails (e.g., light bulb)

$F_X(x) = P(X \leq x)$ $\quad$ $1 - F_X(x) = P(X > x)$

$= P($"system survives" at least to time $x)$

For this reason, $1 - F_X(x)$ is called the survival function $S_X(x) = 1 - F_X(x)$ in medicine and the reliability function $R_X(x) = 1 - F_X(x)$ in engineering.

Earlier we showed that for $X \sim \text{Exponential}(\beta)$ $\Rightarrow$ $F_X(x) = 1 - e^{-\beta x}$ for $x > 0$

So $S_X(x) = R_X(x) = e^{-\beta x}$ for this dist.

The instantaneous failure rate or hazard rate function is defined to be $H_X(x) = \dfrac{f_X(x)}{S_X(x)}$

This gives $\dfrac{P(\text{failure in interval } (x, x+\epsilon) \mid \text{survival to time } x)}{\text{for small } \epsilon > 0} = \dfrac{f_X(x)}{R_X(x)}$

Notice that if $X \sim \text{Exponential}(\beta)$

then $H_X(x) = \dfrac{\beta e^{-\beta x}}{e^{-\beta x}} = \beta \left(\dfrac{\text{Constant in}}{x} \right)$

---

the Exponential is the only failure rate distribution with constant hazard.

Returning to the earlier result that $X \sim \text{Exponential}(\beta)$,

$$\rightarrow P(X \geq t+h \mid X \geq t) = P(X \geq h),$$

for all $t > 0$, $h > 0$

this says that if the product has survived to time t, the chance it will survive to time $(t+h)$ is the same as the original chance of surviving from time 0 to time h; i.e., "the system doesn't remember how long it's survived" (this often makes the Exponential unrealistic in practice)

Consequences ① $X_i \overset{IID}{\sim} \text{Exponential}(\beta)$ $(i=1,\dots,n)$,

then

$Y_1 = \min(X_1, \dots, X_n) \sim \text{Exponential}(n\beta)$.

Beta Distribution $\alpha, \beta > 0$ $X \sim \text{Beta}(\alpha, \beta) \leftrightarrow$

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \cdot \underbrace{I(0<x<1)}_{\text{support of } X}$$

The name comes from the normalizing constant: the function $x^{\alpha-1}(1-x)^{\beta-1}$ has no closed-form anti-derivative, so people just made one

---

Definition For all $\alpha > 0$, $\beta > 0$

$$B(\alpha,\beta) \triangleq \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

B: beta function

Can show that $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

(α, β) jointly control the shape of the $\text{Beta}(\alpha, \beta)$ dist.

$X \sim \text{Beta}(\alpha, \beta)$ | $\psi_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r} \right) \frac{t^k}{k!}$ (yuck)

$E(X) = \dfrac{\alpha}{\alpha + \beta}$ | $V(X) = \left(\dfrac{\alpha}{\alpha+\beta} \right) \left(\dfrac{\beta}{\alpha+\beta} \right) \left(\dfrac{1}{\alpha+\beta+1} \right)$

Case study (Castaneda v. Partida continued): $n = 220$ grand jurors chosen from eligible population of Hidalgo County, Texas, which was 79.1% Mexican-American, but only $s = 100$ selected grand jurors were Mexican-American; let's summarize the <u>information</u> in a Bayesian fashion about evidence of discrimination.

Data S = # Mexican-American chosen in jury selection of $n = 220$ people

Unknown θ = actual probability of an eligible Mexican-American person being chosen ($0 < \theta < 1$)

Sampling Model $(S \mid \theta) \sim \text{Binomial}(n, \theta)$,

i.e., $f_{S|\theta}(s|\theta) = p(S=s|\theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$ (PMF)

$\cdot I(s = 0, 1, \ldots, n)$

Bayesian approach ① Information internal to dataset about θ summarized by the likelihood (un-normalized) density, defined to be $\boxed{p(\theta|s) = c\, p(S=s|\theta)}$,

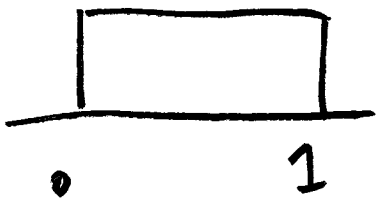
c an arbitrary positive constant — just think of $p(S=s|\theta)$ as a function of θ for fixed s

Here $\ell(\theta|s) = c\binom{n}{s}\theta^s(1-\theta)^{n-s}$ can be absorbed into c since $= c\,\theta^s(1-\theta)^{n-s}$ does not depend on θ

(2) Information external to dataset about θ summarized by the prior density $f_{\Theta}(\theta)$.

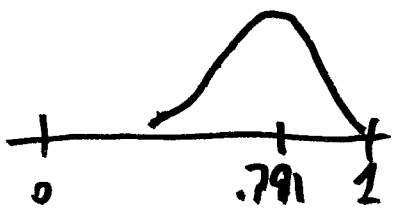
Here are some possibilities for the prior, depending on your knowledge base: (information)

(a) neutral prior $\theta \sim \text{uniform}(0,1)$



this dist. embodies the information {θ could be anywhere between 0 and 1, with no value favored}

(b) Cut the district attorney some slack prior



this prior gives the DA the benefit of the doubt

when you're uncertain about what prior to use, write down all the reasonable priors & do a sensitivity analysis (use each prior essentially one by one & see if posterior answer is the same)

---

③ Combine internal & external information

with Bayes' Theorem

$$f_{\Theta|S}(\theta|s) = c \cdot f_{\Theta}(\theta) \cdot \ell(\theta|s)$$

(posterior information) = (normalizing constant) · (prior information) · (likelihood information)

Here

$$f_{\Theta|S}(\theta|s) = c\, f_{\Theta}(\theta)\, \theta^{s}(1-\theta)^{n-s}$$

Rev. Bayes himself untied back in 1760

that if you take $f_\Theta(\theta) = c\,\theta^{\text{power}_1}(1-\theta)^{\text{power}_2}$ then the product of 2 such densities is another such density, meaning that the posterior would have the same form as the prior & likelihood, making calculating easier

Moreover, we already know the name of densities that look like $\theta^{\text{power}}(1-\theta)^{\text{power}}$: the Beta distributions

$X \sim \text{Beta}(\alpha, \beta)$ $(\alpha > 0, \beta > 0) \rightarrow$

$f_X(x) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$

as our prior PDF

So let's take $f_\Theta(\theta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$ in the lawsuit case study; then

$$f_{\Theta|S}(\theta|s) = c\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]\left[\theta^{s}(1-\theta)^{n-s}\right]$$

$= c\ \theta^{(\alpha+s)-1}(1-\theta)^{(\beta+n-s)-1} = \text{Beta}(\alpha+s, \beta+n-s)$

So the prior-to-posterior updating looks like this:

↑ "Beta dist. is conjugate to the Binomial likelihood"

$\theta \sim \text{Beta}(\alpha, \beta)$

$(s|\theta) \sim \text{Binomial}(n, \theta)$

$\} \rightarrow (\theta|s) \sim \text{Beta}(\alpha+s, \beta+n-s)$

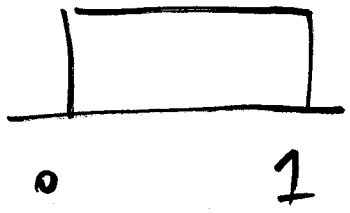
$s = 100$

$n = 220$

How choose (α, β)?

(a) Neutral prior

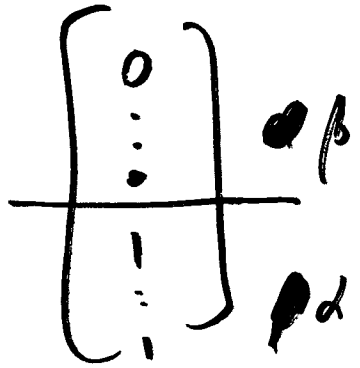
but $\text{Uniform}(0,1) = \theta^{1-1}(1-\theta)^{1-1}$



so $\theta \sim \text{Uniform}(0,1) \leftrightarrow \theta \sim \text{Beta}(1,1)$

(b) cut DA slack prior

There's an extremely useful thing that happens with conjugate priors:

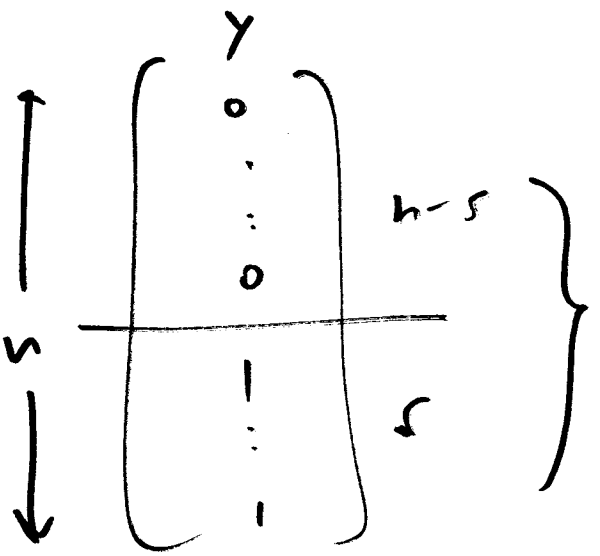


pseudo data

prior effective sample size $(\alpha+\beta)$

mean $\frac{\alpha}{\alpha+\beta}$

Beta prior distribution acts like a dataset with α 1s & β 0s with the property that if you do a Bayesian analysis with the Beta(α,β) prior and I do a frequentist analysis on the dataset with $(\alpha+s)$ 1s and $(\beta+n-s)$ 0s formed by merging the prior & sample data sets, we'll get the same results.



sample data

data set sample size n

mean $\bar{y} = \frac{s}{n}$

(b) cut the DA slack prior

mean of Beta(α, β) dist. is $\frac{\alpha}{\alpha+\beta}$; set this equal to 0.791

Suppose I want to put in prior information equivalent to a prior sample size $\frac{1}{10}$ as big as the data sample size (say); set

$$(\alpha+\beta) = \frac{1}{10} n = 22$$

Solve: $\alpha = 17.4$, $\beta = 4.6$

$n = 220$

$s = 100$

likelihood is

$$c\,\theta^{s}(1-\theta)^{n-s} = c\,\theta^{(s+1)-1}(1-\theta)^{(n-s+1)-1}$$

$$= \text{Beta}(s+1,\ n-s+1) \text{ dist}$$

(101) (121)

(a) Neutral prior: Beta(1,1)

prior sample size = 2

posterior is Beta$(\alpha+s,\ \beta+n-s)$

(101) (121)

(same as likelihood)

(b) Cut DA slack prior: Beta(17.4, 4.6) prior (α, β)

posterior is Beta($\alpha+s$, $\beta+n-s$)

220 → $\alpha+s$ = 117.4; 100 → $\beta+n-s$ = 124.6

prior	posterior mean	posterior SD
neutral	0.455	0.0333
Cut DA slack	0.485	0.0321

Posterior mean of θ is $\frac{\alpha+s}{\alpha+\beta+n}$

Posterior SD is $\sqrt{\left(\frac{\alpha+s}{\alpha+\beta+n}\right)\left(\frac{\beta+n-s}{\alpha+\beta+n}\right)\left(\frac{1}{\alpha+\beta+n+1}\right)}$

The no-discrimination rate of 0.791 is

$$\frac{0.791-0.455}{0.0333} = 10.\bullet$$

$\frac{1}{2}$ posterior SDs away from posterior expectation

under the neutral prior and

$$\frac{0.791 - 0.485}{0.0321} \doteq 9.5 \text{ posterior SDs}$$

away from posterior expectation under the cut-DA-slack prior;

There was
Q.E.D.
discrimination

Multinomial Distributions (back to discrete)

You're contemplating a population that contains elements of $k \geq 2$ types (e.g., {Democrat, Republican, Libertarian, Independent, Green, other}).

Suppose the proportion of elements of type i is $\hat{p}_i$ with $\sum_{i=1}^{k} p_i = 1$; $\underline{p} = (p_1, \ldots, p_k)$.

You take an IID sample of size n from this pop.; X_j = # elements of type i in your sample; $\sum_{i=1}^{k} X_i = n$.

Can show that the vector $\underset{\sim}{X} = (X_1, \ldots, X_k)$ has pmf

$$f_{\underset{\sim}{X}|n,\underset{\sim}{p}}(\underset{\sim}{x} \mid n, \underset{\sim}{p}) = \begin{cases} \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k} & \text{if } \sum_{i=1}^{k} x_i = n \\ 0 & \text{else} \end{cases}$$

$\left(\sum_{i=1}^{k} p_i = 1\right)$

where

$$\binom{n}{x_1, \ldots, x_k} \triangleq \frac{n!}{x_1! x_2! \cdots x_k!}$$ is the multinomial coefficient

---

This is called the multinomial $(n, \underset{\sim}{p})$ distribution.

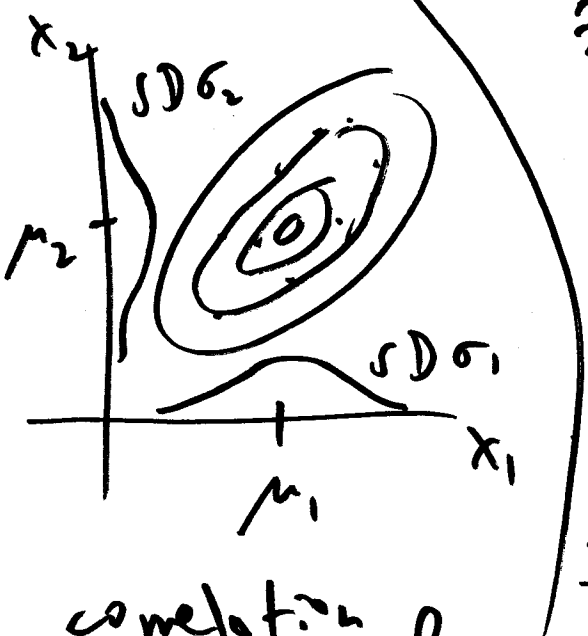
$E(X_i) = np_i$ | $V(X_i) = np_i(1-p_i)$

(just like binomial) | But now something new:

$C(X_i, X_j) = -np_ip_j$ | negatively correlated because $\sum_{i=1}^{k} X_i = n$.

Bivariate Normal Dist. | Can build a 2-dimensional (bivariate) version of the Normal dist. as follows:

$Z_1, Z_2 \overset{IID}{\sim} N(0, 1)$



correlation ρ

Specify 5 parameters:

$-\infty < \mu_1 < +\infty$	$0 < \sigma_1 < \infty$
$-\infty < \mu_2 < +\infty$	$0 < \sigma_2 < \infty$

$-1 < \rho < +1$

Now build $(\underline{X}_1, \underline{X}_2)$ with the transformation $\underline{X}_1 = \mu_1 + \sigma_1 \underline{\zeta}_1$

$$\underline{X}_2 = \sigma_2 \left[\rho \underline{\zeta}_1 + \sqrt{1-\rho^2}\, \underline{\zeta}_2 \right] + \mu_2$$

The joint PDF of $\underline{\underline{X}} = (\underline{X}_1, \underline{X}_2)$ is

then $$f_{\underline{X}_1, \underline{X}_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}\,\sigma_1\sigma_2} \cdot \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right\}$$

standard units

This is the Bivariate Normal $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ dist.

Easy to show that $E(X_1) = \mu_1$,

$E(X_2) = \mu_2$, $V(X_1) = \sigma_1^2$, $V(X_2) = \sigma_2^2$,

$\rho(X_1, X_2) = \rho$. Consequences of this Def.

① $(X_1, X_2) \sim$ Bivariate Normal $\longrightarrow$

$\left(\begin{array}{c} X_1, X_2 \\ \text{independent} \end{array}\right) \longleftrightarrow \left(\begin{array}{c} X_1, X_2 \\ \text{uncorrelated} \end{array}\right)$

we already knew the $\rightarrow$ direction in general; what's new here is that correlation 0 implies independence if $(X_1, X_2) \sim$ Bivariate Normal.

② $(X_1, X_2) \sim$ Bivariate Normal $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$

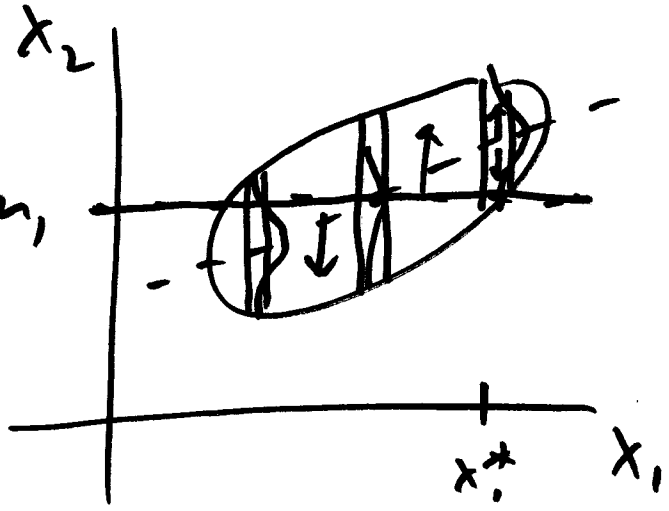
→ conditional distribution of X_2 given that $X_1 = x_1$ is (univariate) normal with mean $E(X_2 \mid x_1) =$

$$\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

and variance $V(X_2 \mid x_1) = (1 - \rho^2)\sigma_2^2$

above result ② says that if (X_1, X_2) are Bivariate Normal then the distributions of X_2 given $X_1 = x_1^*$ in all of the vertical strips are also normal.

Galton, revisited



conditional

And the means of all these normal distributions in the vertical strips are connected together by Galton's

regression line

$$\hat{x}_2 = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1).$$

This line has slope $\beta_1 = \rho\frac{\sigma_2}{\sigma_1}$ and "y"-intercept $\beta_0 = \mu_2 - \beta_1\mu_1$

Moreover, $\hat{x}_2 = \beta_0 + \beta_1 x_1$

we can now quantify an earlier insight:

ignore x_1, predict $(\hat{x}_2)_{\text{no } x_1} = \mu_2 = E(X_2)$

(root mean squared error) (RMSE) of this prediction is

$$\sqrt{V(X_2)} = \sigma_2$$

Use x_1 to predict x_2

$$\text{pred.2t}\,(\hat{x}_2)_{\substack{\text{use}\\ x_1}} = E(X_2 \mid X_1 = x_1)$$

$$= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

RMSE of this prediction is $\sqrt{V(X_2 \mid x_1)} = \sigma_2\sqrt{1-\rho^2}$

~~RMSE(use x_1) $\leq$ RMSE($\hat{x}_{2\,\text{no } x_1}$)~~

Since $-1 < \rho < 1$, $\sigma_2\sqrt{1-\rho^2} \leq \sigma_2$

with equality only when $\rho = 0$.

---

(3) $(X_1, X_2) \sim$ BivariateNormal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$,

$Y = a_1X_1 + a_2X_2 + b$, (a_1, a_2, b) arbitrary constants

$\rightarrow Y \sim N(a_1\mu_1 + a_2\mu_2 + b,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2\rho\sigma_1\sigma_2)$.

Large Random Samples

(DS ch. 6)

You draw an IID random sample $X_1, \ldots, X_n$ from a population, with the goal of estimating the population mean $\mu = E(X_i)$.

We've already seen that, from a root mean squared error point of view, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the best you can do (in the absence of prior information).

It would be nice if $\bar{X}_n$ approached the right answer μ as n increases; how to quantify that idea?

Two inequality that help

Markov inequality

Suppose X is a non-negative r.v., i.e. $P(X \geq 0) = 1$ then for all real $t > 0$, $P(X \geq t) \leq \frac{E(X)}{t}$ ⊛

(Attributed to Andrey Markov (1856-1922), a Russian mathematician who did pioneering work on stochastic processes)

⊛ says that, if $E(X)$ is fixed, you can't move more & more probability out into the right tail beyond a certain point.

25 Aug 11

Laplace
↕
Bertels → Lobachevsky → Brashman → Chebyshev → Markov → Voronoy → Sierpinski → Neyman → Lehmann → Draper