

Def | rv $X_1, \dots, X_n \rightarrow$ sample mean

of (X_1, \dots, X_n) is $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

Consequence,
continued

(5) $\left\{ \begin{array}{l} X_i \stackrel{IID}{\sim} N(\mu, \sigma^2) \\ (i=1, \dots, n) \end{array} \right\}$

$\rightarrow \bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

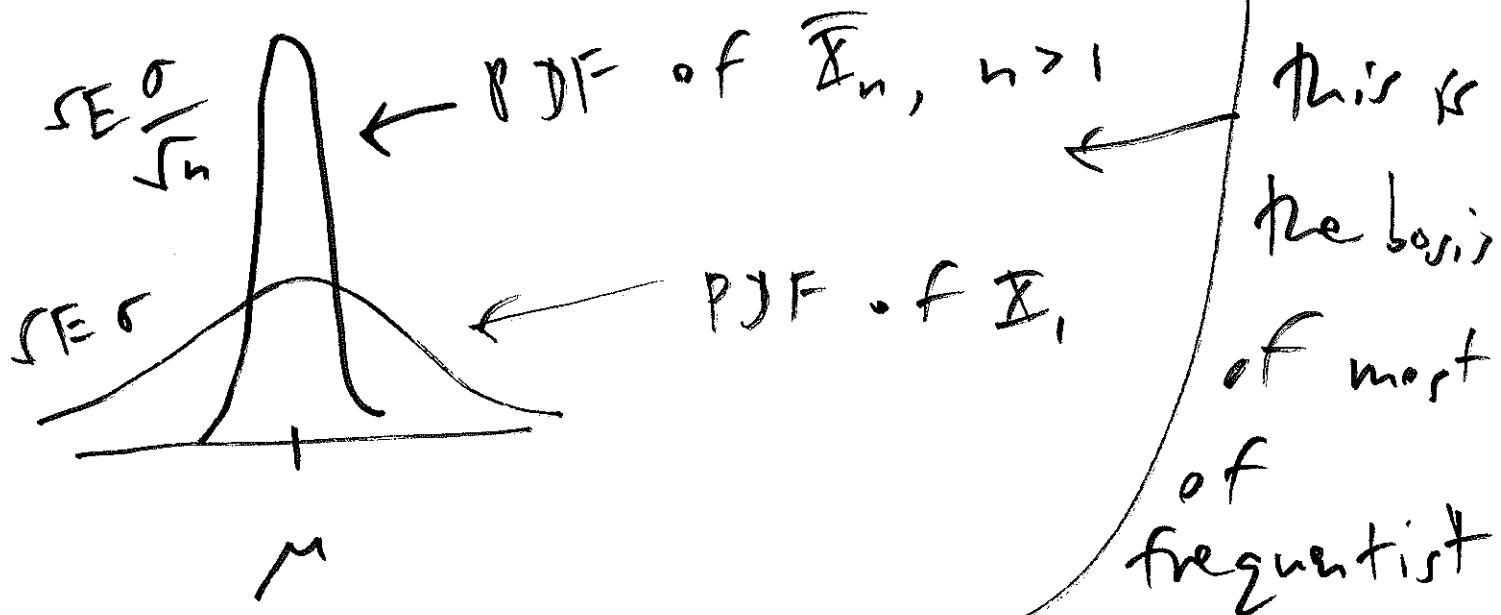
so $SD(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$

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 $\hat{\theta}_n^{(rv)}$ of a parameter θ is called the standard error $SE(\hat{\theta})$ of $\hat{\theta}_n$

So if you use \bar{X}_n as an estimate ⁽²⁶⁷⁾ of μ , $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \rightarrow 0$ as $n \rightarrow \infty$



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, you have to quadruple the sample size.

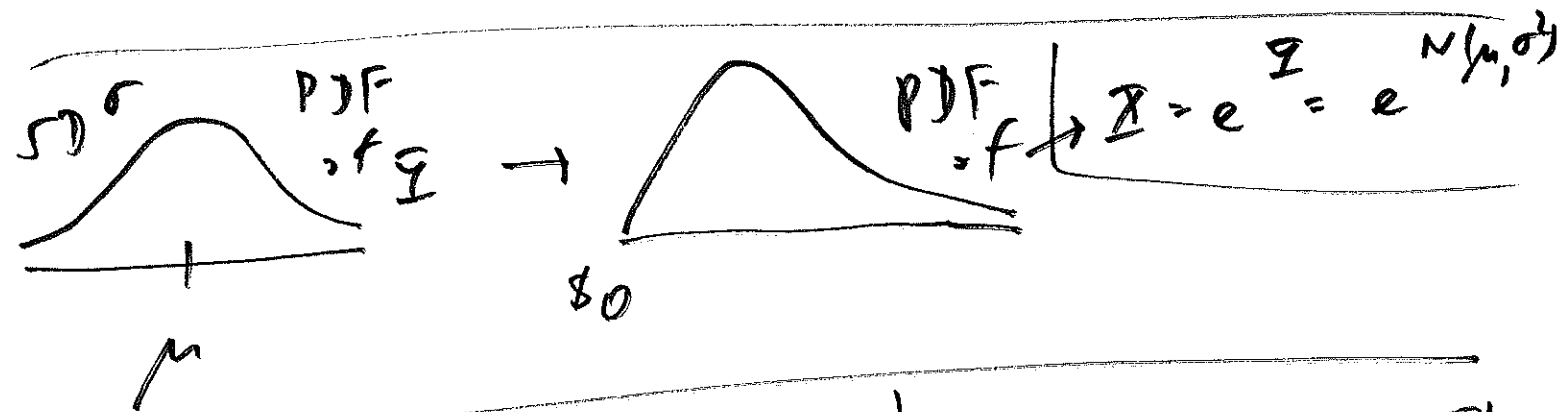
log normal Distribution (This distribution is mis-named; it should be called the

Exponential-Normal distribution, but we're stuck with a bad name.)

Def.

$X > 0$

If $Z = \log(X) \sim N(\mu, \sigma^2)$, people say that $X \sim$ Log Normal (μ, σ^2) .



$X \sim$ LogNormal (μ, σ^2)

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

~~scribbled out text~~

Can get MGF of X from MGF of Z

MGF of \mathbb{I} is $\psi_{\mathbb{I}}(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$ (269)

But by definition

$$\psi_{\mathbb{I}}(t) = E(e^{t\mathbb{I}}) = E(e^{t \log X})$$

$$= E(X^t), \text{ so we can}$$

$$E(X) = \psi_{\mathbb{I}}(1)$$

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

read the moments of X directly from the MGF of \mathbb{I}

$$V(X) = \psi_{\mathbb{I}}(2) - (\psi_{\mathbb{I}}(1))^2$$

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

Famous case study

~~example~~

(known constant)

price S_0 . Heroic assumption: price

Pricing stock options, continued

1 share of a stock, current

u time units in the future will be 270

$$S'_u = S'_0 e^{\xi_{1u}}, \quad \xi_{1u} \sim N(\mu u, \sigma^2 u).$$

Can write $S'_0 e^{\xi_{1u}} = e^{\xi_{1u} + \log(S'_0)}$. Now

$$\left[\xi_{1u} + \log(S'_0) \right] \sim N(\mu u + \log(S'_0), \sigma^2 u),$$

$$\text{So } S'_u \sim \text{Log Normal}(\mu u + \log(S'_0), \sigma^2 u).$$

Consider a single time horizon u ;

heroic
assumption
rewritten \rightarrow

$$S'_u = S'_0 \exp[\mu u + (\sigma\sqrt{u}) \cdot \xi_1],$$

$$\xi_1 \sim N(0, 1)$$

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

Use risk-neutral pricing as in the (271) previous discussion: force present value

$$E(S_u) \stackrel{\Delta}{=} S_0.$$

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

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

But by heroic ^{log normal} assumption,

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

So set S_0 equal to

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

for risk-neutral pricing.

Value of option at time u will be (272)

$$h(S_u), \text{ where } h(s) = \begin{cases} s - g & \text{if } s > g \\ 0 & \text{else} \end{cases}$$

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

$$\frac{1}{2} > \frac{\log\left(\frac{g}{S_0}\right) - \left(r - \frac{\sigma^2}{2}\right)u}{\sigma\sqrt{u}} \triangleq c$$

Now a nasty integral

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

which is

$$e^{-ru} E[h(S_u)] = e^{-ru} \int_c^{\infty} \left[S_0 e^{(r - \frac{\sigma^2}{2})u + \sigma z \sqrt{u}} - g \right]$$

Careful calculation reveals the (famous) formula $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$

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

the risk-neutral price of the option,

where $c = \log\left(\frac{q}{S_0}\right) - \left(r - \frac{\sigma^2}{2}\right)u$ ← This formula

(Black-Scholes formula)

was derived in 1973 by

Gamma Distribution

(American economist)

Fischer Black (1938-1995)

($\alpha, \beta > 0$) I has the

Gamma dist. with parameters (α, β),

Canadian-American economist

and Myron Scholes (1941-)

with $I \sim \Gamma(\alpha, \beta)$ or

$I \sim \text{Gamma}(\alpha, \beta) \rightarrow$

won Nobel prize in Economics for this work in 1997, together with Robert

I continuous on $(0, \infty)$ with

American (economist)

Merton (1944-2003)

PDF $f_{\mathbb{R}}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{I}(x > 0)$

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,

involved to deal with integrals of functions like $\textcircled{*}$ above:

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

has no anti-derivative in closed form

(275)

$\Gamma(x)$ turns out to be a continuous generalization of the factorial function,

because $\left(\begin{array}{c} n \text{ positive} \\ \text{integer} \end{array} \right) \rightarrow \Gamma(n) = (n-1)!$

$\Gamma(x) \rightarrow \infty$ really quickly as $x \rightarrow \infty$,
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
handle $\Gamma(x)$ is with a Stirling's

approximation: $\Gamma(x) \approx \sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x}$
for large x

so that $\ln f(x) = \frac{1}{2} \ln(2\pi) + (x - \frac{1}{2}) / \ln x - x$ (276)

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

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.

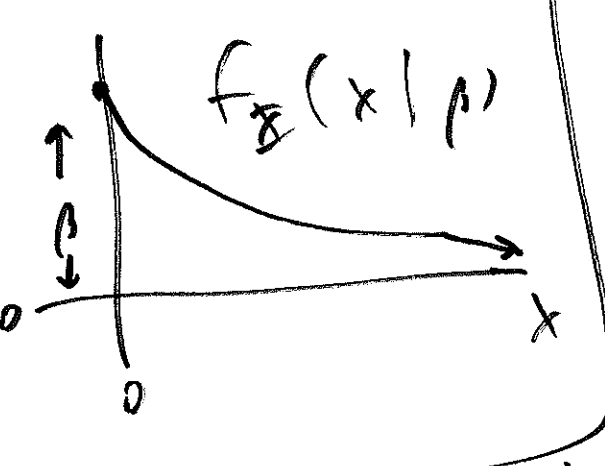
$X \sim \text{Exponential}(\beta)$

$$f_X(t) = \frac{\beta}{\beta - t}, \quad t < \beta$$

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

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

$$D(X) = \frac{1}{\beta}$$



~~Notice that the Exponential distribution has $D(X)$ equal to $E(X)$; this suggests it's related somehow to the Poisson dist.~~

The over

Suppose that arrivals (events) occur according to a Poisson process with rate β per unit time.

and define $T_1 = T_1 - 0$

$$T_2 = T_2 - T_1$$

$$\dots T_k = T_k - T_{k-1} \text{ for } k = 2, 3, \dots$$

Set $T_k =$ time until k^{th} arrival
 $k = 1, 2, \dots$

The T_i are called the inter-arrival (278)

times.

Then it turns out that $T_i \stackrel{\text{IFD}}{\sim} \text{Exponential}(\beta)$

The

Exponential dist. is also related to the Geometric dist., in that they both

have a memoryless property Theorem

$X \sim \text{Exponential}(\beta)$; $t > 0$, $h > 0$

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

Example) $X =$ ^{from initial use} time until a manufactured product fails (eg., light bulb)

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

$= P(\text{"system surviving" at least to time } x)$

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 $F_X(x) = 1 - e^{-\beta x}$
for $X \sim \text{Exponential}(\beta)$ 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 $H_X(x) = \frac{f_X(x)}{S_X(x)}$

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

Notice that if $X \sim \text{Exponential}(\beta)$ (280)

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

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
 $t \geq 0$
 $h \geq 0$

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)

Consequence ① $X_i \stackrel{i.i.d.}{\sim}$ Exponential (β) (281)
 $(i=1, \dots, n)$,

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

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

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

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

Definition For all $\alpha > 0$
 $\beta > 0$ $B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$
 \uparrow
beta function

Can show that $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. (282)

(α, β) jointly control

the shape of the Beta (α, β) dist.

(yuck)

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

$E(X) = \frac{\alpha}{\alpha+\beta}$

$V(X) = \left(\frac{\alpha}{\alpha+\beta} \right) \left(\frac{\beta}{\alpha+\beta} \right) \left(\frac{1}{\alpha+\beta+1} \right)$

Case Study

~~DeLoe~~

(Castaneda v. Partida continued)

$n=220$ grand jurors chosen from ~~(eligible)~~ eligible population of Hidalgo County, Texas, which was 79.1% Mexican-American, but only $s=100$

selected grand jurors were Mexican-American; summarize the information in a Bayesian fashion about evidence of discrimination.

Data $S = \#$ Mexican-American ^{chosen} in jury selection of $n = 220$ people (283)

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

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

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

Bayesian approach ① Information internal to data set about θ summarized

by the likelihood (un-normalized) density, defined to be $\ell(\theta | S) = c P(S=s|\theta)$,

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

Here $L(\theta | s) = c \binom{4}{s} \theta^s (1-\theta)^{4-s}$ can be absorbed into c since c does not depend on θ

$$= c \theta^s (1-\theta)^{4-s}$$

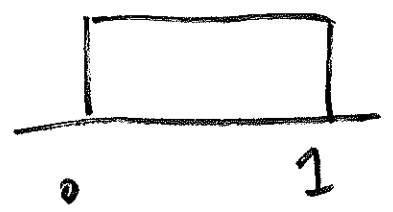
(2) Information external to dataset about θ summarized by the prior density $f(\theta)$.

Here are some

possibilities for the prior, depending on your knowledge base:

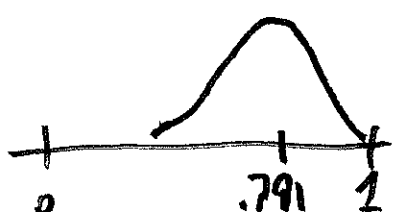
your knowledge base:

(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) but the district attorney somewhat prior



this prior gives the DA the benefit of the doubt

when you're uncertain about what prior 285
 to use, write down all the reasonable priors
 & do a sensitivity analysis (use each prior
 one by one & see if ^{posterior} answer is the same) essentially

③ Combine internal & external information

with
 Bayes'
 Theorem

$$f_{\theta|S}(\theta|S) = c \cdot f_{\theta}(\theta) \cdot L(\theta|S)$$

\uparrow \uparrow \uparrow
 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 noticed back in 1760

that if you take $f_{\theta}(\theta) = c \theta^{\text{power}} (1-\theta)^{\text{power}}$ 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 calculations

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

the $X \sim \text{Beta}(\alpha, \beta)$ ($\alpha > 0, \beta > 0$) \rightarrow Beta distributions $f_X(x) = c \theta^{\alpha-1} (1-\theta)^{\beta-1}$

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

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

$$= c \theta^{(\alpha+s)-1} (1-\theta)^{(\beta+n-s)-1} = \text{Beta}(\alpha+s, \beta+n-s) \quad (287)$$

So the prior-to-posterior

updating looks like this:

Beta dist. is conjugate to the Binomial likelihood

$$\left. \begin{array}{l} \theta \sim \text{Beta}(\alpha, \beta) \\ (s' | \theta) \sim \text{Binomial}(n, \theta) \end{array} \right\} \rightarrow (\theta | s) \sim \text{Beta}(\alpha+s, \beta+n-s)$$

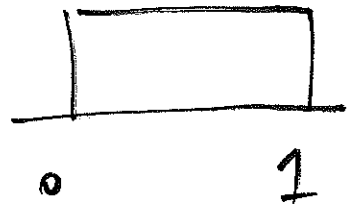
$$s = 100$$

$$n = 220$$

How choose (α, β) ?

(a) Neutral prior

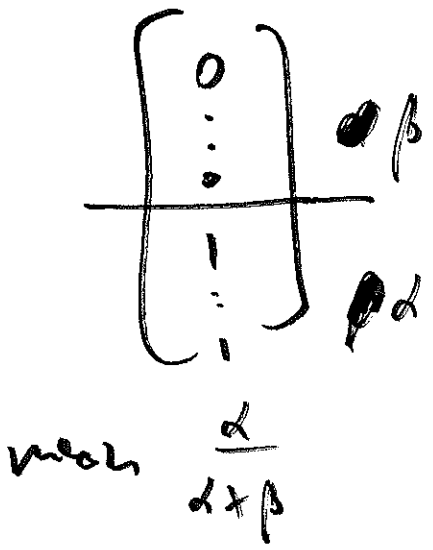
$$\text{but Uniform}(0, 1) = \theta^{1-1} (1-\theta)^{1-1}$$



$$\text{So } \theta \sim \text{Uniform}(0, 1) \Leftrightarrow \theta \sim \text{Beta}(1, 1)$$

(b) cut
DA
stock
prior

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

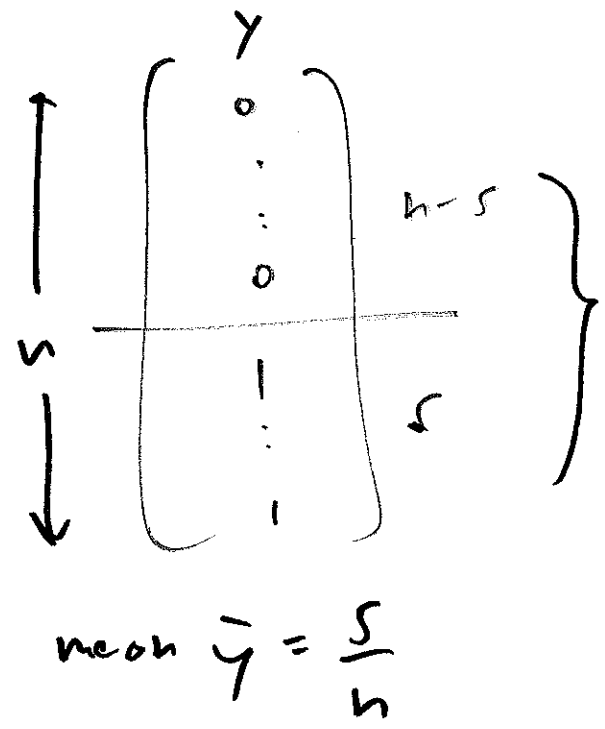


pseudo data

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

Beta prior distribution acts like a dataset with α 1s & β 0s

with the property that



sample data

dataset sample size n

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.

(b) Cut the DA stock prior

mean of Beta(α, β) dist. is $\frac{\alpha}{\alpha + \beta}$ (289)

$\frac{\alpha}{\alpha + \beta}$; 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 (507); set

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

Solve: $\begin{cases} \alpha = 17.4 \\ \beta = 4.6 \end{cases}$

$$n = 220$$

$$s = 100$$

likelihood is

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

= Beta($s+1, n-s+1$) dist

(101)

(121)

(a) Neutral prior:

$$\text{Beta}(1, 1)$$

prior is

$$\text{Beta}(\alpha + s, \beta + n - s)$$

\uparrow
101

\uparrow
(121)

prior sample size 2

(same as likelihood)

(b) Cut DA stock prior

Beta (α, β) prior

290

posterior \rightarrow Beta ($\alpha + s, \beta + n - s$)
 \uparrow \uparrow
 117.4 124.6

220 100
 \downarrow \downarrow

prior	posterior		posterior mean of θ is $\frac{\alpha + s}{\alpha + \beta + n}$
	mean	SD	
neutral	0.455	0.0333	
cut DA stock	0.485	0.0321	

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

$\frac{0.791 - 0.455}{0.0333} = 10.2$ posterior SDs away from posterior expectation

under the neutral prior and

(291)

$$\frac{0.791 - 0.485}{0.0321} = 9.5 \text{ posterior S.D.s}$$

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}).

Suppose the proportion
of elements of type i is $0 \leq p_i \leq 1$
with $\sum_{i=1}^k p_i = 1$; $\mathbf{p} = (p_1, \dots, p_k)$.

You take an IID sample of size n from this pop.; $X_i = \#$ elements of type i in your sample; $\sum_{i=1}^k X_i = n$. (282)

Can show that the vector $\underline{X} = (X_1, \dots, X_k)$

has M.P.F. $f_{\underline{X}|n, \underline{p}}(x_1, \dots, x_k) = \begin{cases} \binom{n}{x_1, \dots, x_k} p_1^{x_1} \dots p_k^{x_k} & \text{if } \sum_{i=1}^k x_i = n \\ 0 & \text{else} \end{cases}$

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

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

This is called the Multinomial (n, \underline{p}) distribution.

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

(just like binomial)

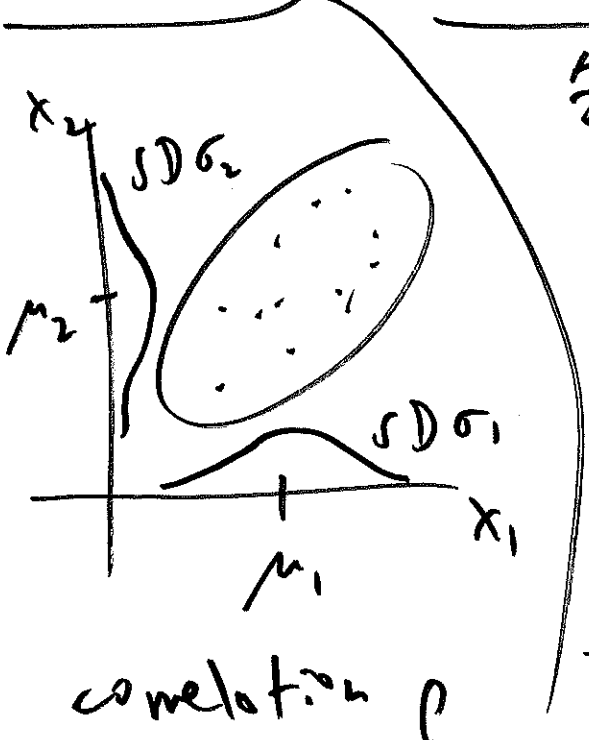
But now something new:

$C(X_i, X_j) = -n p_i p_j$

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:



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

Specify 5 parameters:

$-\infty < \mu_1 < +\infty$	$0 < \sigma_1 < \infty$
$-\infty < \mu_2 < +\infty$	$0 < \sigma_2 < \infty$
$-1 < \rho < +1$	

Now build (X_1, X_2) with the transformation $X_1 = \mu_1 + \sigma_1 Z_1$

$$X_2 = \sigma_2 \left[\rho Z_1 + \sqrt{1-\rho^2} Z_2 \right] + \mu_2$$

The joint PDF of $\underline{X} = (X_1, X_2)$ is

$$\text{then } f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_1\sigma_2} \cdot \exp \left\{ \right.$$

$$-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right.$$

$$\left. \left. + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

standard units

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

Easy to show that $E(X_1) = \mu_1$, (295)

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

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

① $(X_1, X_2) \sim \text{Bivariate Normal} \rightarrow$

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

we already knew the \rightarrow direction is general; what's new here is that correlation 0 implies independence

if $(X_1, X_2) \sim \text{Bivariate Normal}$.

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

\rightarrow conditional distribution of X_2

given that $X_1 = x_1$ is (univariate)

normal with mean $E(X_2 | x_1) =$

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

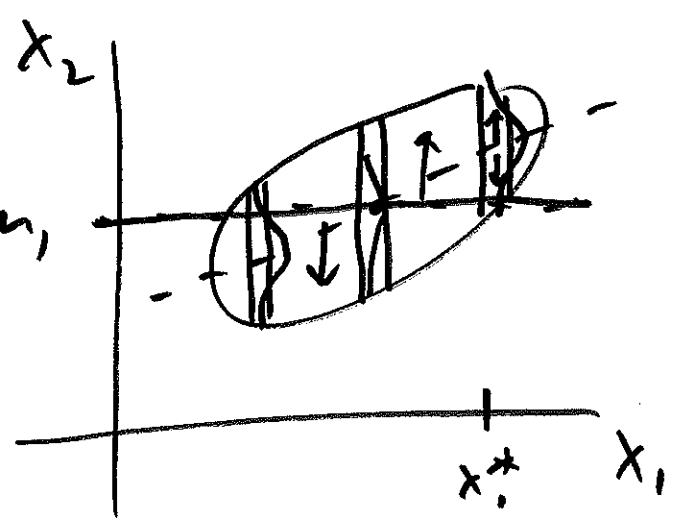
and variance $V(X_2 | x_1)$

$$= (1 - \rho^2) \sigma_2^2$$

above

Result ② says that if (X_1, X_2) are

Galton, revisited



conditional

Bivariate Normal then the distributions of X_2 given $X_1 = x_1^*$ in all of the vertical strips are also normal

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

regression line

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

This line has slope $\beta_1 = \frac{\rho\sigma_2}{\sigma_1}$ 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 ,

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

(root mean squared error)

(RMSE) of this prediction is

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

use x_1
to predict
 x_2

$$\text{pred. of } (\hat{x}_2)_{\text{use } x_1} = E(X_2 | X_1 = x_1)$$

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

RMSE of this

prediction is $\sqrt{V(X_2 | x_1)} = \sigma_2 \sqrt{1 - \rho^2}$

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

with equality only when $\rho = 0$.

③ $(X_1, X_2) \sim \text{Bivariate Normal}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$

$Y = a_1 X_1 + a_2 X_2 + b$, (a_1, a_2, b) arbitrary constants

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

Large
Random
Samples

(DS ch. 6)

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

We've already seen that, from a worst case point of view, the best 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 inequalities that help

Markov inequality

Suppose

X is a non-negative r.v., i.e.

$$P(X \geq 0) = 1$$

Then for all

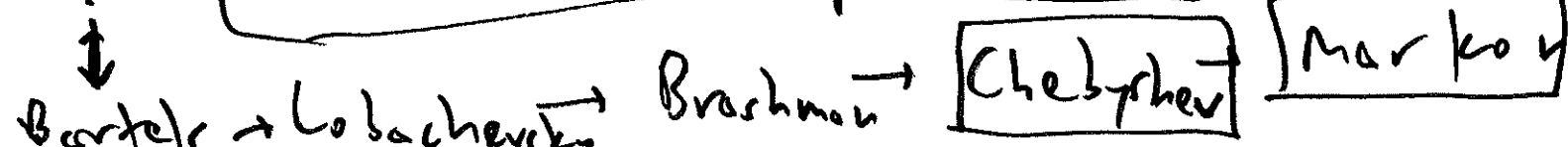
$$\text{real } t > 0, \quad P(X \geq t) \leq \frac{E(X)}{t} \quad *$$

(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.

Laplace

25 April



ex. | $E(X) = 1$, X non-negative \rightarrow

$$P(X \geq 100) \leq \frac{1}{100}$$

The inequality is

sharp, meaning that the upper bound

$\frac{E(X)}{t}$ on $P(X \geq t)$ is attainable, \otimes

ex. | $E(X) = 1$, X -nonnegative \rightarrow

put probability 0.99 on $X=0$ and
0.01 on $X=100$

\otimes but most of the time (i.e., for most distributions) it's a crude upper bound.

Can apply Markov inequality to the

v.v. $Y = [X - E(X)]^2$ to get

Chebyshev Inequality } X r.v. with $V(X)$ existing (302)
→ for every $t \geq 0$,

$$P\left(|X - E(X)| \geq t\right) \leq \frac{V(X)}{t^2}$$

(attributed to

Pafnuty Chebyshev (1821 - 1894), also a Russian mathematician, one of whose Ph.D. students was Markov)

Ex.

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

Chebyshev says $P\left[\left|\frac{X - \mu}{\sigma}\right| \geq 3\right] \leq \frac{1}{3^2} = \frac{1}{9}$,

so no more than $\frac{1}{9} = 11\%$ of the probability in any distribution, with finite variance, can

be more than 3 SDs away from the mean (recall for Normal dist. this prob. is 0.3%)

This upper bound is also sharp, but (3.3)
for most distributions it's (also) crude
(as with the Markov bound). Back to \bar{X}_n

$X_i \stackrel{iid}{\sim}$ some dist. with mean $E(X_i) = \mu$
($i=1, \dots, n$) and variance $V(X_i) = \sigma^2 < \infty$

Then we already showed that if $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

then $E(\bar{X}_n) = \mu$ for all $n=1, 2, \dots$
and $V(\bar{X}_n) = \frac{\sigma^2}{n}$.

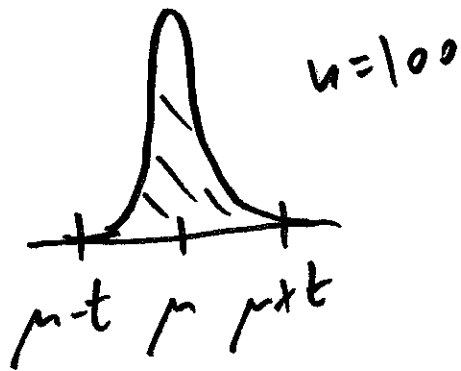
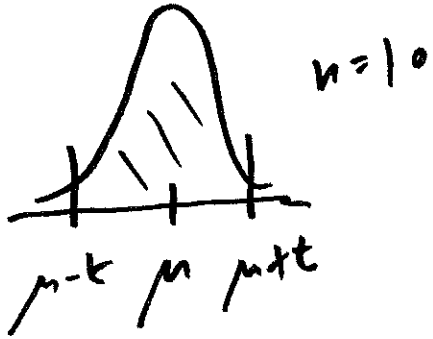
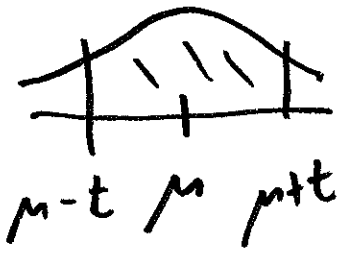
Chebyshev then

$$\text{gives } P(|\bar{X}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2} \text{ for all } t > 0$$

this can be

rewritten
$$P(|\bar{X}_n - \mu| < t) \geq 1 - \frac{\sigma^2}{nt^2}$$

PDF of \bar{X}_n $n=1$



⋮

This suggests a way ⁽³⁰⁴⁾ to quantify how close a r.v. like \bar{X}_n is to a constant like μ :

Def. A sequence Z_1, Z_2, \dots of r.v. is said to converge in probability to a constant b if

for all $\epsilon > 0$, $\lim_{n \rightarrow \infty} P(|Z_n - b| < \epsilon) = 1$;

this is denoted $Z_n \xrightarrow{P} b$.

An immediate

consequence of Chebyshev's & this definition is