5/30/19 Lecture 18

* There will be an extra lecture next Wednesday evening through Webcast → more details soon

<u>Def</u>: rv $X_1, \dots, X_n$ → sample mean of $(X_1, \dots, X_n)$ is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Consequences continued

5) $X_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$ $(i = 1, \dots, n)$ (All of them)

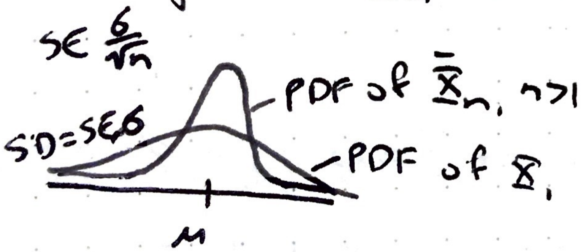
$\rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (Each of them)

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 $\hat{\theta}(n)$ 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, $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$



This is the basis of most frequentist statistical inference

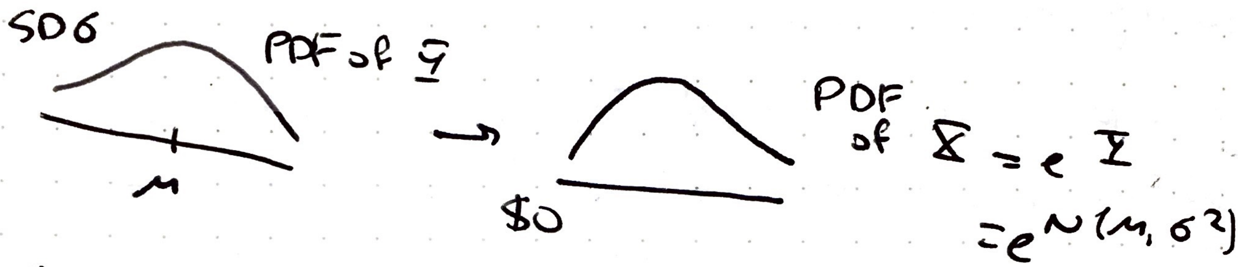
As $n \uparrow$, $\bar{X}_n$ gets better as an estimate of μ, at a $\frac{1}{\sqrt{n}}$ rate. This is called the Square Root Law.

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

Lognormal Distribution

should be exponential normal

Def: If $X > 0$ and $Y = \log(X) \sim N(\mu, \sigma^2)$, then $X \sim \text{Lognormal}(\mu, \sigma^2)$



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

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

Can get MGF of X from MGF of Y

MGF of Y is $\psi_Y(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$

But by def: $\psi_Y(t) = E(e^{tY}) = E(e^{t \log X}) = E(X^t)$

So we can read the moments of X directly from the MGF of Y

$E(X) = \psi_Y(1) = \exp(\mu + \frac{\sigma^2}{2})$

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

Gamma distribution

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

→ X continuous on $(0, \infty)$ with...

PDF $X \sim \Gamma(\alpha, \beta)$

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

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 distribution is

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

$$\Gamma(\alpha) \triangleq \int_0^\infty \underbrace{x^{\alpha-1} e^{-x}}_{\text{has no anti-derivative in closed form}} dx$$

$\Gamma(\alpha)$ turns out to be a continuous generalization of the factorial function, because (n positive integer)

→ $\Gamma(n) = (n-1)!$

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

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

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

$X \sim \Gamma(\alpha, \beta)$

$\psi_X(t) = (1 - \frac{t}{\beta})^{-\alpha}$ for $t < \beta$

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

Alternate expression

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

Special case of $\Gamma(\alpha, \beta)$

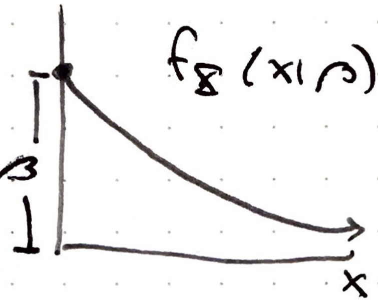
With $\alpha = 1$ the PDF is $f_X(x \mid \beta) = \beta e^{-\beta x} I(x > 0)$

This is just the exponential distribution

$X \sim$ Exponential (β)

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

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



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

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

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

and define $Y_1 = Z_1 - 0$

$Y_2 = Z_2 - Z_1$

$\dots \quad Y_k = Z_k - Z_{k-1}$ for $k = 2, 3, \dots$

Set Z_k = time until kth arrival $k = 1, 2, \dots$

The Y_i are the inter-arrival times.

$Y_i \overset{IID}{\sim}$ Exponential (β) as a result.

The exponential dist. is also related to the Geometric dist. in that they both have a memoryless property.

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

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

Ex: X = time from initial use until a manufactured product fails

(e.g. light bulb)

$F_X(x) = P(X \le x)$

$1 - F_X(x) = P(X > x)$

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

For this reason, $1 - F_X(x)$ is 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

For $X \sim$ 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 $H_X(x) = \dfrac{f_X(x)}{S_X(x)} = \dfrac{f_X(x)}{R_X(x)}$

This gives $P($failure in interval $(x, x+\epsilon) \mid$ survival to time $x)$ for small ϵ

Notice that if $X \sim$ Exponential(β) then $H_X(x) = \dfrac{\beta e^{-\beta x}}{e^{-\beta x}}$ $= \beta$ (constant in x)

The exponential is the only failure rate distribution with constant hazard.

Returning to the earlier result that $X \sim$ Exponential(β)

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

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

"the system doesn't remember how long it survived"
(This often makes the Exponential unrealistic in practice)

Consequences

1) $X_i \overset{iid}{\sim}$ Exponential (β) $(i=1,\dots,n)$ then
$Y_1 = \min(X_1,\dots,X_n) \sim$ Exponential$(n\beta)$

Beta Distribution

$\alpha, \beta > 0$

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

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 a definition...

Def: for all $\alpha>0$ $\beta>0$

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

↑
beta function

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

(α, β) jointly control the shape of the Beta(α, β) dist.

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

$$\psi_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r} \right) \cdot \frac{t^k}{k!}$$

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

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

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 $0 < p_i < 1$ with $\sum_{i=1}^{k} p_i = 1$ $\quad \underset{\sim}{p} = (p_1, \dots, p_k)$

You take an IID sample of size n from this pop.

X_i = # elements of type i in your sample

$$\sum_{i=1}^{k} X_i = n$$

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

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

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

This is called the Multinomial (n, p) distribution

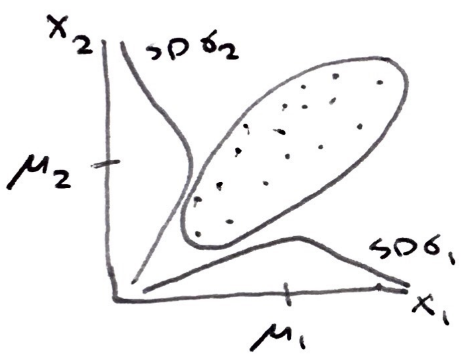
$$\left.\begin{aligned} E(X_i) &= np_i \\ V(X_i) &= np_i(1-p_i) \end{aligned}\right\} \text{ just like binomial}$$

$$C(X_i, X_j) = -np_i p_j$$

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:



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

specify 5 parameters

1) $-\infty < \mu_1 < \infty$
2) $-\infty < \mu_2 < \infty$
3) $0 < \sigma_1 < \infty$
4) $0 < \sigma_2 < \infty$
5) $-1 < \rho < 1$

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 $\underset{\sim}{X} = (X_1, X_2)$ is then

$$f_{X_1 X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}\,\sigma_1\sigma_2} \cdot \exp\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

standard units

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

Easy to show that $E(X_1) = \mu_1$, $E(X_2) = \mu_2$

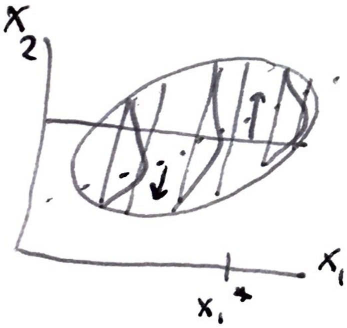
$V(X_1) = \sigma_1^2$, $V(X_2) = \sigma_2^2$, $\rho(X_1, X_2) = \rho$

Consequences of this def:

1) $(X_1, X_2) \sim$ Bivariate Normal $\rightarrow \begin{pmatrix} X_1, X_2 \\ \text{independent} \end{pmatrix} \leftrightarrow \begin{pmatrix} X_1, X_2 \\ \text{uncorrelated} \end{pmatrix}$

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 $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \rightarrow$ conditional distribution of X_2 given that $X_1 = x_1$ is (univariate) normal with mean $E(X_2 | x_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ and variance $V(X_2 | x_1) = (1-\rho^2)\sigma_2^2$



Galton revisited

Result 2 says that if (X_1, X_2) are Bivariate Normal then the conditional distributions of X_2 given $X = x_1^*$ 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

$$\hat{x}_2 = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) \qquad \hat{x}_2 = \beta_0 + \beta_1 x_1$$

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

Moreover, we can now quantify an earlier insight

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

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

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

$\boxed{\text{use } x_1 \text{ to predict } x_2}$ predict $(\hat{x}_2)_{\substack{use \\ x_1}} = E(X_2 \mid X_1 = x_1)$

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

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

Since $-1 < \rho < 1$, $\sigma_2 \sqrt{1-\rho^2} \le \sigma_2$ with equality only when $\rho = 0$

3) $(X_1, X_2) \sim$ 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

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

Large random samples

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 inequalities that help

Markov inequality

Suppose X is a non-negative rv, ie $P(X \geq 0) = 1$

Then for all real $t > 0$, $P(X \geq t) \leq \frac{E(X)}{t}$ says that if $E(X)$ is fixed, you can't move more and more probability out into the right tail beyond a certain point

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 but most of the time it's a crude upper bound

Ex: $E(X) = 1$, X - non-negative $\rightarrow$ put probability 0.99 on $X = 0$ and probability 0.01 on $X = 100$