Part 1 10.3

Last lecture, we defined a special r.v. $T \sim \exp(\lambda)$ if density is $f(t) = \lambda e^{-\lambda t}$, $t > 0$ (λ is a positive constant)

$$F(t) = P(T \le t) = 1 - e^{-\lambda t}$$

$$E(T) = \int_{-\infty}^{\infty} t f(t)\,dt = \int_0^{\infty} t \cdot \lambda e^{-\lambda t}\,dt$$

$$= \cancel{t \cdot \cancel{\lambda} \frac{e^{-\lambda t}}{-\cancel{\lambda}}\Big|_0^{\infty}} \to 0 - \int_0^{\infty} 1 \cdot \cancel{\lambda} \frac{e^{-\lambda t}}{-\cancel{\lambda}}\,dt$$

$$E(T) = \frac{1}{\lambda}$$

$$Var(T) = E(T^2) - (E(T))^2 = E(T^2) - \frac{1}{\lambda^2}$$

$$E(T^2) = \int_0^{\infty} t^2 \cdot f(t)\,dt = \int_0^{\infty} t^2 \cdot \lambda e^{-\lambda t}\,dt = \frac{2}{\lambda^2}$$

(Exercise)

$$Var(T) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}, \quad SD(T) = \frac{1}{\lambda}$$

$T \sim \exp(\lambda)$, $f(t) = \lambda e^{-\lambda t}$, $t > 0$ (λ is positive constant)

$F(t) = 1 - e^{-\lambda t}$, $t > 0$ ($F(t) = 0$, $t \le 0$)

$$E(T) = \frac{1}{\lambda}$$

$$Var(T) = \frac{1}{\lambda^2}$$

$$SD(T) = \frac{1}{\lambda}$$

Memoryless property of T

We often use the exponential dsn to model lifetimes that are random. (machine parts, lightbulbs) We want to see how long something will last.

Let $T \sim exp(\lambda)$, $P(T \leq t) = F(t) = 1 - e^{-\lambda t}$

$$P(T > t) = e^{-\lambda t}$$

$$S(t) := P(T > t) = e^{-\lambda t}$$ survival function

$$P(T > s+t \mid T > t) = \frac{P(T > s+t \ \& \ T > t)}{P(T > t)}$$

$$= \frac{P(T > s+t)}{P(T > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}}$$

$$P(T > s+t \mid T > t) = \underline{P(T > s)}$$

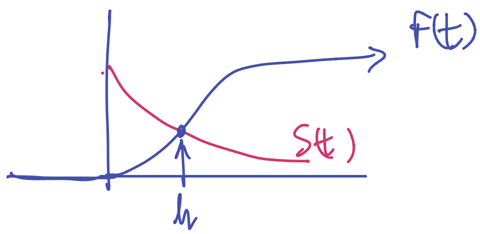
So, given that the lifetime is longer than t, the chance that the lifetime will be more than s time units past t is just the chance that the lifetime will be longer than s time units if it started from scratch. This is called the memoryless property of the exponential dsn.

<u>Note</u> we cannot use this dsn to model the lifetimes of things that age. P(human will live past 100 years | Human age > 50) $\neq$ P(human lives past 50)

Median of the exponential dsn

· Median is the point where half the dsn is to left & half the dsn to the right

$$P(T \leq t) = P(T < t) = P(T > t)$$



$$S(h) = F(h)$$
$$e^{-\lambda h} = 1 - e^{-\lambda h}$$
$$e^{-\lambda h} = \frac{1}{2}$$

h is the median of the dsn (halfway point)

$$e^{-\lambda h} = \frac{1}{2} \rightarrow e^{\lambda h} = 2$$
$$\lambda h = \log(2) \quad (\log_e \; \ln)$$
$$h = \frac{\log(2)}{\lambda} = \log(2) \cdot E(T)$$

Note that $\log(2) \approx 0.69 < 1$

Median(T) $<$ E(T) $\Rightarrow$ Distribution is skewed right

---

Half-life : We use exp. dsn to model lifetimes of radioactive atoms that decay at an exponential rate

Half-life : time until half the atoms have decayed. of a radioactive isotope

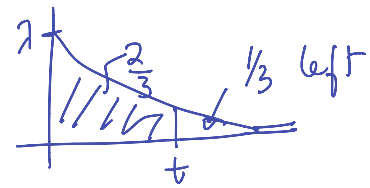
λ is called "decay rate"

$$h = \frac{\log(2)}{\lambda} \Rightarrow \lambda = \frac{\log(2)}{h}$$

Example (10.5.4) Strontium-90 has a half-life of 28.8 years. Assume exponential decay, how long will it take for $\frac{2}{3}$ of a lump of Strontium-90 to decay.

$$h = 28.8 = \frac{\log(2)}{\lambda}, \text{ so } \lambda = \frac{\log(2)}{28.8} \approx 0.024$$

Now we want t s.t.



$P(T > t) = \frac{1}{3}$ $\quad (P(T \leq t) = \frac{2}{3})$

$$F(t) = \frac{2}{3} \Rightarrow 1 - e^{-\lambda t} = \frac{2}{3}, \quad \lambda = 0.024$$

$$e^{-\lambda t} = \frac{1}{3} \Rightarrow \lambda t = \log(3)$$

$$t = \frac{\log(3)}{0.024} \approx 45.78 \text{ years.}$$

This sort of computation is used in "Carbon dating" (radiocarbon dating) which uses the proportion of Carbon-14 (^{14}C) found in organic matter to determine its age.

Say we know rate of decay λ for ^{14}C

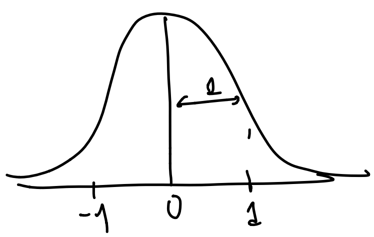
Look at proportion of ${}^{14}C$ left ($P(T>t)$)

$$p = e^{-\lambda t} \Rightarrow e^{\lambda t} = \frac{1}{p}$$

$$t = \frac{1}{\lambda}\log\left(\frac{1}{p}\right)$$

Part 2 10.4. The Normal Distribution

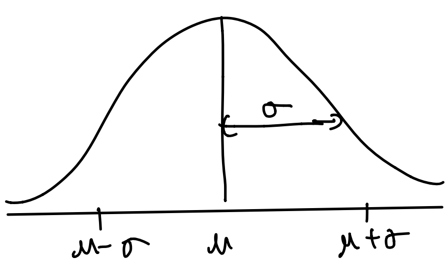
This is a well known curve $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$



We can define a more general density function that is centered at any $\mu \in \mathbb{R}$, spread is $\sigma > 0$

We say that the r.v. X has the normal distribution $X \sim N(\mu, \sigma^2)$ if the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty \text{ all real } x$$



If $\mu = 0, \sigma = 1$ we call this r.v. Z and say Z has the standard normal curve

We usually denote the density of Z by φ (instead of f) and the cdf of Z by Φ

$$\Phi(z) = P(Z \le z)$$

If $X \sim N(\mu, \sigma^2)$, $Z \sim N(0,1)$

Z is "z-score of X" or X in standard units

$$X = \sigma Z + \mu, \quad Z = \frac{X-\mu}{\sigma}$$

$\mathbb{E}(Z) = 0$, $Var(Z) = 1$, $SD(Z) = 1$

$\mathbb{E}(X) = \mu$, $Var(X) = \sigma^2$, $SD(X) = \sigma$

<u>Important fact</u>:

If X, Y are independed normally distributed r.v.

$X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

This extends to other linear combinations

say $X - 3Y + 5 \sim N(\mu_X - 3\mu_Y + 5, \sigma_X^2 + 9\sigma_Y^2)$

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

We use this to find C.I. for the difference of means.

Say n, m are sample sizes, large enough for the CLT to apply

$X_1, X_2, \ldots X_n \sim$ iid $E(X_k) = \mu_X$, $SD(X_k) = \sigma_X$

$Y_1, Y_2, \ldots Y_m \sim$ iid, $E(Y_k) = \mu_Y$, $SD(Y_k) = \sigma_Y$

$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$, $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$

So $\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$

approx. 95% C.I. for $\mu_X - \mu_Y$ is given by

$$\left.\begin{aligned} &\bar{X} - \bar{Y} \pm 2 \cdot SD(\bar{X} - \bar{Y}) \\ &= \bar{X} - \bar{Y} \pm 2 \cdot \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \end{aligned}\right\}$$

list of "plausible" values for $\mu_X - \mu_Y$

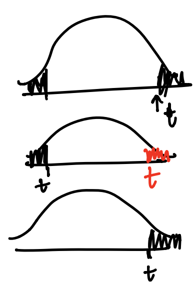
We can set up hypothesis tests for equality of means

$H_0: \mu_X = \mu_Y$

H_1 or H_A: $\mu_X \neq \mu_Y$ $\quad T = |\bar{X} - \bar{Y}|$

$\mu_X < \mu_Y$ $\quad T = \bar{Y} - \bar{X}$

$\mu_X > \mu_Y$ $\quad T = \bar{X} - \bar{Y}$



large values support the alternative

<u>Note</u> we are only making a statement about $\mu_X - \mu_Y$ not the entire distribution.

A special case is when μ_X, μ_Y are proportions.
This is when X_i's Y_i's are Bernoulli

$X_1, \ldots, X_n \sim \text{Bernoulli}(p_X)$

$Y_1, \ldots, Y_m \sim \text{Bernoulli}(p_Y)$

Then $\underbrace{\bar{X} - \bar{Y}}_{\text{fractions}} \sim N\left(p_X - p_Y, \frac{p_X q_X}{n} + \frac{p_Y q_Y}{m}\right)$

We can use the normal dsn to find C.I.
& hypothesis tests.

<u>Difference for hypothesis tests</u>

$H_0: p_X = p_Y \leftarrow \quad T = \bar{X} - \bar{Y}$ (one sided test)

$H_1: p_X > p_Y$

$\bar{X} - \bar{Y} \overset{\text{under } H_0}{\sim} N\left(0, \frac{pq}{n} + \frac{pq}{m}\right)$ $\quad p =$ common value p_X & p_Y

To estimate p, combine both samples into one big sample & count # of successes total.

$\bar{X} = \frac{\text{\# of successes}}{n} = \frac{\sum X_i}{n}$, $\quad \bar{Y} = \frac{\sum Y_i}{m}$

$n\bar{X} + m\bar{Y} =$ total # of successes

$\hat{p} = \frac{n\bar{X} + m\bar{Y}}{n+m}$ $\leftarrow$ total # of successes $\sum X_i + \sum Y_i$
$\leftarrow$ total sample size