LECTURE 25 : THURSDAY APRIL 21

Part 1 : Finish §10.2, expectation & variance

Part 2 : Two special distributions

~~Part 3 : Bias and variance~~

~~Part 4 : The German Tank problem~~

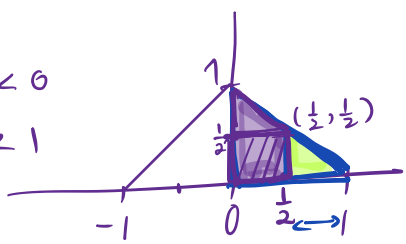
} we will do two on Tuesday

PART 1

$f(x) \geq 0$ s.t $\int_{-\infty}^{\infty} f(x)\,dx = 1$ → density function
under density curve
areas describe probs for a continuous r.v.

Let X be a r.v. with the density function $f(x)$ described below:

$$f(x) = \begin{cases} 0, & x < -1 \\ x+1, & -1 \leq x < 0 \\ 1-x, & 0 \leq x < 1 \\ 0, & x \geq 1 \end{cases}$$

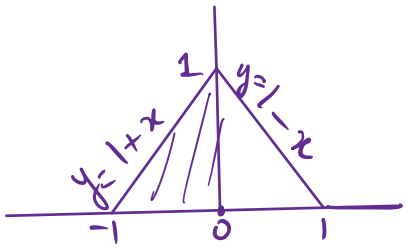


Find ① $P\left(\frac{1}{2} < X < 1\right) = \frac{1}{2} b \cdot h = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$

② $P\left(0 < X < \frac{1}{2}\right) = \frac{1}{2} - \frac{1}{8} = \frac{4-1}{8} = \frac{3}{8}$
$= \text{area of sq} + \text{area } \triangle = \left(\frac{1}{2}\right)^2 + \frac{1}{8} = \frac{3}{8}$

$F(x) =$ Area under the curve up to x

$= \int_{-\infty}^{x} f(t)\,dt = P(X \leq x)$



$$F(x)=\begin{cases}0, & x<-1\\ \int_{-1}^{x}(1+t)\,dt, & -1\le x<0\\ \frac{1}{2}+\int_{0}^{x}(1-t)\,dt, & 0\le x\le 1\\ 1, & x>1\end{cases}$$

$$\int_{-1}^{x}(1+t)\,dt = t+\frac{t^2}{2}\Big|_{-1}^{x}$$

$$= x+\frac{x^2}{2}-\left(-1+\frac{1}{2}\right) = x+\frac{x^2}{2}+\frac{1}{2} = \frac{x^2+2x+1}{2}$$

$$\int_{0}^{x}(1-t)\,dt = t-\frac{t^2}{2}\Big|_{0}^{x} = x-\frac{x^2}{2} = \frac{2x-x^2}{2}$$

$$F(x)=\begin{cases}0, & x<-1\\ \frac{x^2+2x+1}{2}, & -1\le x\le 0\\ \frac{1}{2}+\frac{2x-x^2}{2}=\frac{1+2x-x^2}{2}, & 0\le x\le 1\\ 1, & x>1\end{cases}$$

$$P\left(\tfrac{1}{2}<X<1\right) = F(1)-F\left(\tfrac{1}{2}\right)$$

$$= \frac{1+2-1}{2} - \overbrace{\frac{1+2\left(\frac{1}{2}\right)-\frac{1}{4}}{2}}^{2-1/4=7/4}$$

$$= 1-\frac{7}{8}=\frac{1}{8}$$

EXPECTED VALUE & VARIANCE

- $P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$

Note $f(x) \neq P(X=x) \neq 0$ for continuous r.v. X.

- $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \mu$
- $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx.$
- $\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$

$g(X) = (X-\mu)^2$, $\mathrm{Var}(X) = \mathbb{E}(g(X))$

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx$$

$$= \int_{-\infty}^{\infty} [x^2 - 2\mu x + \mu^2] f(x)\,dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\,dx - 2\mu \underbrace{\int_{-\infty}^{\infty} x f(x)\,dx}_{=\mu} + \mu^2$$

$$\underbrace{\qquad\qquad}_{-2\mu^2} \qquad + \mu^2$$

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - \left(\int_{-\infty}^{\infty} x f(x)\,dx\right)^2$$

PROPERTIES

$\int cf(x) = c\int f(x)$

(1) $\mathbb{E}(aX+b) = a\mathbb{E}(X)+b$

(2) $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$

$SD(aX+b) = |a|\,SD(X)$

(3) $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$

(4) If for every interval A, B

$P(X\in A, Y\in B) = P(X\in A)\,P(Y\in B)$

then we say that X & Y are independent r.v.

(5) If X, Y are independent r.v.

then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

$\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

PART 2 : TWO SPECIAL DISTRIBUTIONS

- The Exponential distribution

$$f(x) = e^{-x}, \quad x > 0 \quad (0 \text{ elsewhere})$$

- The r.v. X that has the pdf $f(x) = e^{-x}$, $x > 0$ is called an EXPONENTIAL r.v. WITH RATE 1

$$X \sim \exp(1)$$

- In general, we say $X \sim \exp(\lambda)$ if $f(x) = \lambda e^{-\lambda x}$, $x > 0$, $\lambda > 0$ (λ is some constant)

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{0}^{x} \lambda e^{-\lambda t}\,dt$$

$$= \lambda \int_{0}^{x} e^{-\lambda t}\,dt = \lambda \cdot \frac{-e^{-\lambda t}}{\lambda}\Bigg|_{0}^{x} = 1 - e^{-\lambda x}$$

$$F(x) = P(X \le x) = 1 - e^{-\lambda x}, \quad x > 0$$

$$(F(x) = 0, \quad x \le 0)$$

We often denote the exponential r.v. by T.

$$T \sim \exp(\lambda), \quad P(T \le x) = 1 - e^{-\lambda x}, \quad \begin{matrix} x > 0 \\ \lambda > 0 \end{matrix}$$

$$P(T > x) = 1 - P(T \le x)$$
$$= 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$$

MEMORYLESS PROPERTY OF $T \sim \exp(\lambda)$

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

$$\mathbb{E}(T) = \frac{1}{\lambda} \leftarrow \text{use integration by parts}$$

$$\mathbb{E}(T) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x}\,dx$$

$$\mathbb{E}(T) = \underbrace{x \cdot \frac{\lambda e^{-\lambda x}}{-\lambda}\Big|_0^{\infty}}_{0} + \int_0^{\infty} \frac{\lambda e^{-\lambda x}}{+\lambda}\,dx$$

$$\mathbb{E}(T) = \int_0^{\infty} e^{-\lambda x}\,dx = \frac{e^{-\lambda x}}{-\lambda}\Big|_0^{\infty} = \frac{1}{\lambda}$$

$$\mathrm{Var}(T) = \mathbb{E}(T^2) - (\mathbb{E}(T))^2$$

$$= \underbrace{\mathbb{E}(T^2)}_{\frac{2}{\lambda^2}} - \frac{1}{\lambda^2}$$

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

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

$$F(t) = \begin{cases} 1 - e^{-\lambda t}, & t > 0 \\ 0, & t \leq 0 \end{cases}$$

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

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

$P(T > t) = S(t)$, S is the survival function

T models (randomly distributed) lifetime

S ... is the chance of lasting past time t

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

This means that the chance that the object will get past s more time units if it has already lasted t time units is the same as the object lasting s time units when new

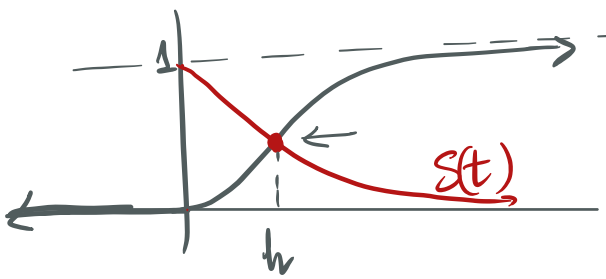
Object forgets that it has lasted for t time units.

Median of the exponential distribution:

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

$$P(T \le t) = P(T < t) = \underbrace{P(T > t)}_{e^{-\lambda t}}$$

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



$$S(t) = P(T > t) = 1 - F(t)$$

at h: $S(h) = F(h)$

$$e^{-\lambda h} = 1 - e^{-\lambda h}$$

$$2e^{-\lambda h} = 1$$

$$e^{-\lambda h} = \frac{1}{2}$$

h is the median or halfway point of the dsn.

$$e^{-\lambda h} = \frac{1}{2}$$

$$e^{\lambda h} = 2$$

$$\lambda h = \ln(2) \iff \log(2)$$

$$h = \frac{\log(2)}{\lambda} = \left(\frac{\ln(2)}{\lambda}\right) = \ln(2) \cdot \mathbb{E}(T)$$

$\ln(2) \approx 0.69 < 1$

Median $< \mathbb{E}(T)$ $\rightarrow$ Dsn is skewed right

HALF-LIFE

The half-life of a radioactive isotope is the time until half the atoms have decayed.

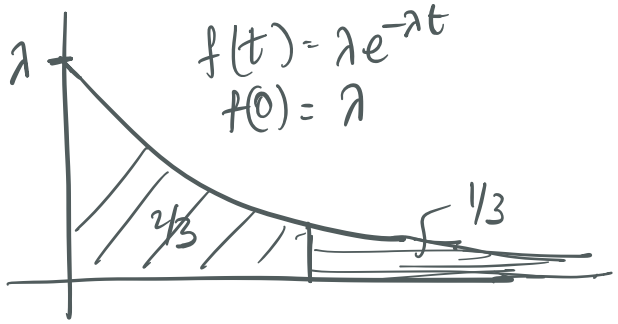
λ is called the decay rate & we use $T \sim \exp(\lambda)$ to model lifetime of radioactive atom

Ex 10.5.4 Strontium-90 which has a half life of 28.8 years.
If we assume exponential decay, how long will it take for $\frac{2}{3}$ of a lump of Strontium-90 to decay?

Half-life h is value s.t. $P(T<h) = P(T>h)$

h is given, $h = 28.8$ years $= \frac{\log(2)}{\lambda}$

$$\lambda = \frac{\log(2)}{h} \approx \frac{0.69}{28.8} \approx 0.024$$



$f(t) = \lambda e^{-\lambda t}$
$f(0) = \lambda$

Want t s.t.
$P(T>t) = \frac{1}{3}$

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

$\Rightarrow t = \frac{\log(3)}{\lambda} \approx 45.78$ years.

This is the kind of computation in Carbon-dating which uses the proportion of Carbon 14 in organic matter to determine its age.

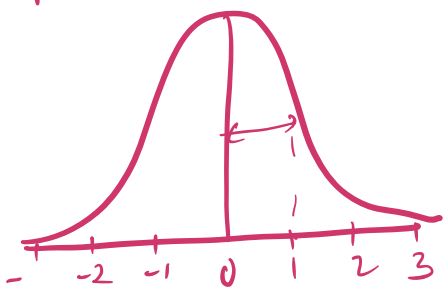
Look at $P(T > t)$ = prop. of carbon 14 left.

$$p = e^{-\lambda t}$$

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

THE NORMAL DISTRIBUTION §10.4

Well-known density curve $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$ $(-\infty, \infty)$



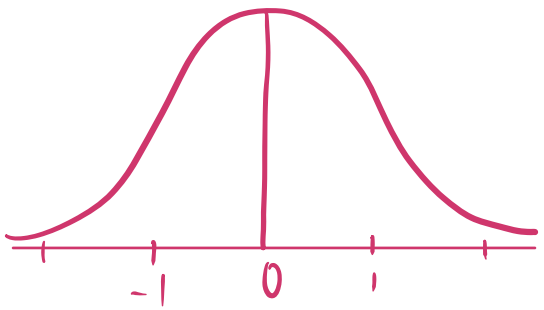
We can define a more general density function centered at $\mu \in \mathbb{R}$, spread $\sigma > 0$

We say that X has the Normal dsn $N(\mu, \sigma^2)$ if the density of X is given by

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

If $\mu=0$, $\sigma^2=1$ ($\sigma=1$), then we denote the associated r.v. by Z and call this the Standard Normal dsn.

$$F(x) = \Phi(x) = P(Z \le x)$$

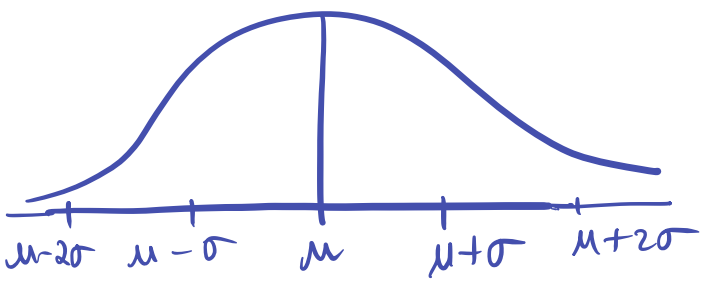


$Z \sim N(0,1)$

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

If we put X in standard units

$$Z = \frac{X-\mu}{\sigma}$$



$$X = \sigma Z + \mu$$

$$E(Z) = 0, \quad Var(Z) = 1, \quad SD(Z) = 1$$

$$E(X) = \mu, \quad Var(X) = \sigma^2, \quad SD(X) = \sigma$$

IMPORTANT FACT

If X, Y are independent & normally distributed

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

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

For instance

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

↑ why?

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

This r.v. is what we use to compute ## confidence intervals for the difference in means.

Suppose we have $X_1, X_2, \ldots X_n \sim$ iid μ_X, σ_X^2

$Y_1, Y_2, \ldots Y_m$, μ_Y, σ_Y^2

By CLT for n, m large enough.

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right), \quad \overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$$

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

Approx 95% CI for $\mu_X - \mu_Y$

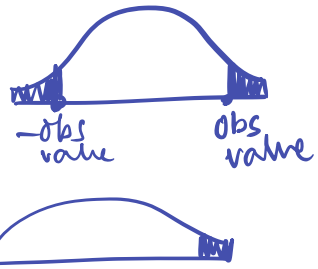
$$(\overline{X} - \overline{Y}) \pm 2 \times SD(\overline{X} - \overline{Y})$$

$$: (\overline{X} - \overline{Y}) \pm 2 \times \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

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

Setting up hyp. tests for equality of means

Test	Statistic
$H_0: \mu_X = \mu_Y$	
$H_A: \mu_X \neq \mu_Y$	$\lvert T \rvert = \lvert \bar{X} - \bar{Y} \rvert$
$\mu_X < \mu_Y$	$T = \bar{Y} - \bar{X}$
$\mu_X > \mu_Y$	$T = \bar{X} - \bar{Y}$



You are making a statement about $\mu_X - \mu_Y$

Proportions C.I for true difference in proportions

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

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

$$\bar{X} - \bar{Y} \underset{\text{CLT}}{\approx} N\left(p_X - p_Y, \frac{p_X q_X}{n} + \frac{p_Y q_Y}{m}\right)$$

Difference for hypothesis tests.

$H_0: p_X = p_Y$

$H_A: p_X > p_Y$ (Suppose we are doing a one-sided test)

$T = \bar{X} - \bar{Y}$

Under assumption H_0 is true, $p_X = p_Y = p$ ← common value

To estimate p, we combine or POOL our samples.

X : [0] [0] [0] [1] ... [1] ....

Y : [0] [0] [1] [0] [1] ...

# of successes in X sample = avg (total size)

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

X_i, Y_i are 0-1 tkt values

$$\sum X_i = n\bar{X}, \quad \sum Y_i = m\bar{Y}$$

Total # of [1]'s in combined sample $= n\bar{X} + m\bar{Y}$

$\hat{p}$ = estimate of p from the combined sample

$$= \frac{n\bar{X} + m\bar{Y}}{m+n}$$

$n\bar{X} + m\bar{Y}$ ← total # of successes or tkts marked [1]

$m+n$ ← sample size

Example 2 indep samples from Fresno ($n=400$) & Irvine ($m=500$)

proportion in Fresno sample that agree with lifting mask mandate on airplanes: 57%

" Irvine " " : 61%

$\bar{X} = 0.57$, $\bar{Y} = 0.61$ $\quad \sigma_{\bar{X}}^2 \approx \frac{\hat{p}_X(1-\hat{p}_X)}{n} = \frac{(0.57)(0.43)}{400}$

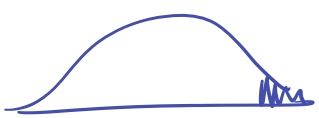
$\sigma_{\bar{Y}}^2 = \frac{\hat{p}_Y(1-\hat{p}_Y)}{m} = \frac{(0.61)(0.39)}{500}$

$SD(\bar{Y}-\bar{X}) = \sqrt{\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2}$

$(\bar{Y}-\bar{X}) \pm 2 \cdot SD(\bar{Y}-\bar{X})$

$H_0: p_X = p_Y$

$H_1: p_X < p_Y$ $\quad T = \bar{Y} - \bar{X}$



$\bar{X} = 0.57$

$n\bar{X} = (0.57)(400)$, $(n\bar{Y}) = (0.61)(500)$

$\hat{p} = \frac{(0.57)(400)+(0.61)(500)}{400+500} \approx 0.5922$

$\sigma_{\bar{Y}-\bar{X}} = \sigma = \sqrt{\sigma_{\bar{X}} + \sigma_{\bar{Y}}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{400} + \frac{\hat{p}(1-\hat{p})}{500}}$

$Z = \frac{(\bar{Y}-\bar{X})-0}{\sigma}$