- Please bring a calculator
- Bring a crib sheet (front & back of an 8.5" x 11" paper)
- Bring Midterm 2 topics pdf

Review for Midterm 2

Continue to build on

- Basics of probability
- Conditional probability
- Distributions of discrete R.V.'s.
  - Bernoulli
  - Binomial
  - Geom
  - Geom*

New material

- Poisson dist (discrete R.V.) $X \sim \text{Pois}(\lambda)$

  - PMF $P(X=k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$

    - Fixed time window of consideration Δt
    - λ = the expected # of events in that time window

  - CDF $F_X(x) = \sum_{j \le x} \left(\dfrac{e^{-\lambda}\lambda^k}{k!} \right)$

- $E(X) = \lambda$
- $Var(X) = \lambda$
- Let Y denote the inter-arrival time, $Y \sim exp(\lambda)$

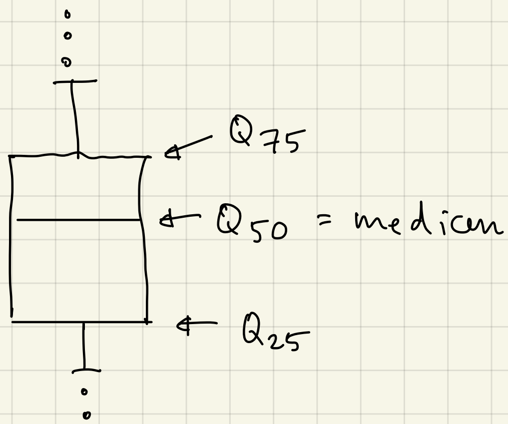
  written as # of events per unit time

- Superposition of two independent Poisson processes

  If $X_1 \sim Pois(\lambda_1)$ and $X_2 \sim Pois(\lambda_2)$

  $$X_1 + X_2 \sim Pois(\lambda_1 + \lambda_2)$$

- Boxplots and quantiles
  - the n-th quantile, denoted Q_n, is the value in the dataset for which n percent of the other values are below it.



"whiskers" extend upwards from Q_{75} to $Q_{75} + \text{range} \cdot IQR$

extend downwards from Q_{25} to $Q_{25} - \text{range} \cdot IQR$

$IQR = Q_{75} - Q_{25}$ the Inter-Quartile-Range

- Points beyond the whiskers are outliers

- Continuous R.V., Y

Sample space $\quad S = \{-\infty < Y < \infty\}$

- PDF requires
$$\begin{cases} f_Y(x) \geq 0 \\ \int_{-\infty}^{\infty} f_Y(x)\,dx = 1 \end{cases}$$

- CDF $\quad F_Y(k) = P(Y \leq k) = \int_{-\infty}^{k} f_Y(x)\,dx$

- $E(Y) = \int_{-\infty}^{\infty} x\, f_Y(x)\,dx$

- $Var(Y) = \underbrace{E(Y^2)}_{= \int_{-\infty}^{\infty} x^2 f_Y(x)\,dx} - [E(Y)]^2$

Recall $\quad P(Y = x) = 0$

$$P(x \leq Y \leq x + dx) = f_Y(x)\,dx$$

e.g. $\quad P(Y = 1.01) = 0$

$$P(1.01 \leq Y \leq 1.01 + dx) = f_Y(1.01)\,dx$$

Many times we encountered a continuous R.V. and then made that into a Bernoulli trial. Then we assessed properties of a collection of Bernoulli trials.

e.g. HW4, # 2.2 ; HW5 # 2.3

- Common distributions for continuous R.V.s.

  - <u>Uniform dist</u>, $Y \sim unif(a, b)$

    - PDF $$f_Y(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

    - CDF $$F_Y(k) = P(Y \leq k) = \frac{k-a}{b-a}$$

    - $E(Y) = (b+a)/2$

    - $Var(Y) = (b-a)^2/12$

  - <u>Exponential dist</u>, $Y \sim Expo(\lambda)$

    - PDF $$f_Y(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- CDF $F_Y(k) = P(Y \le k) = 1 - e^{-\lambda k}$

This means $P(Y > k) = 1 - P(Y \le k) = e^{-\lambda k}$

- $E(Y) = 1/\lambda$

- $Var(Y) = 1/\lambda^2$

- λ is the number of events expected per unit time

(In a Poisson dist, λ is # of events expected in a specified time window Δt).

- Describes the duration time
  - service time for a CPU
  - waiting time between events.
    (e.g. earthquakes, bus arrivals)

- <u>Normal / Gaussian dist</u>

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

- PDF $f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ for $-\infty < x < \infty$

- CDF $F_Y(k) = \int_{-\infty}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx$

- $E(Y) = \mu$

- $Var(Y) = \sigma^2$

- Standard normal distribution

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

- PDF $\quad f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

- CDF $\quad F_Z(k) = P(Z \leq k) \equiv \Phi(k)$

tabulated in a "z-table" for positive z values

$$\Phi(-z) = 1 - \Phi(z)$$

Can translate any $N(\mu, \sigma^2)$ into $N(0, 1)$

If $Y \sim N(\mu, \sigma^2)$ then $\frac{Y-\mu}{\sigma} \sim N(0,1)$

e.g. $P(Y \leq k) = P\left(\frac{Y-\mu}{\sigma} \leq \frac{k-\mu}{\sigma}\right) = \Phi\left(\frac{k-\mu}{\sigma}\right)$

* Can use Z-tables to calculate CDF for any normally distributed R.V., given z find $\Phi(z)$

* Can also use the z-table to find quantiles.

e.g. Q_{25} is the value of z for which $\Phi(z) = 0.25$

Q_{75} is the value of z for which $\Phi(z) = 0.75$

. Binary classification HW5

The sample can have $D=0$, $D=1$.

The test result (the classifier) gives $T=0$, $T=1$.

- Prevalence $\pi = P(D=1)$

- Sensitivity (TPR) $= P(T=1 \mid D=1) = \dfrac{P(D=1 \mid T=1)\, P(T=1)}{P(D=1)}$

- Specifity (TNR) $= P(T=0 \mid D=0) = \dfrac{P(D=0 \mid T=0)\, P(T=0)}{P(D=0)}$