Chapter 5: Continuous Probability Distributions

- A continuous random variable "takes on decimal values"
- For such random variables, the probability of any specific value is 0.

Eg.) What is Pr(a man is exactly 6', 2.357921761 23")?.
Pr(a man is exactly 6'2") ≈ 0
↳ Men are a little above, a little below

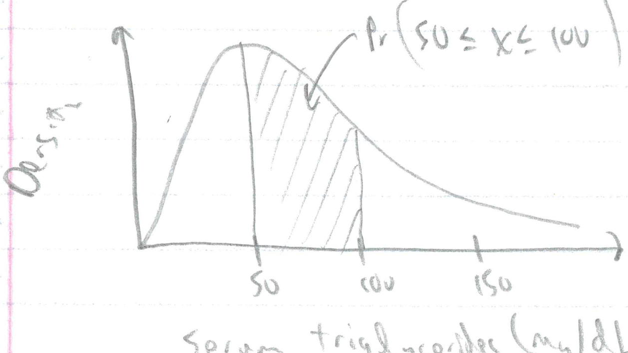
- But, we know some regions are more likely than others!

$$Pr(5' \le X \le 7') > Pr(0' \le X \le 1')$$

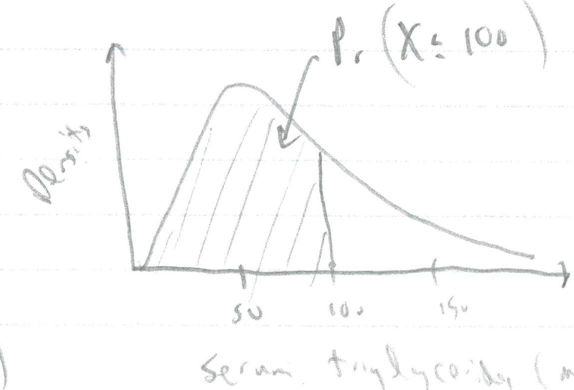
- We describe this intuition with a PDF
- A Probability Density Function of a RV X is a function s.t.

$$Pr(a \le x \le b) = \text{area below curve between a and b}$$

- The CDF is again the $F(x) = Pr(X \le x)$
- Ex) X = Serum triglyceride level



CDF



- Expected value: μ, average x over many trials

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{where } f(x) = \text{density}$$

- Variance: Average squared distance

$$\sigma^2 = E\left((x-\mu)^2\right) = E(x^2) - \mu^2$$

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

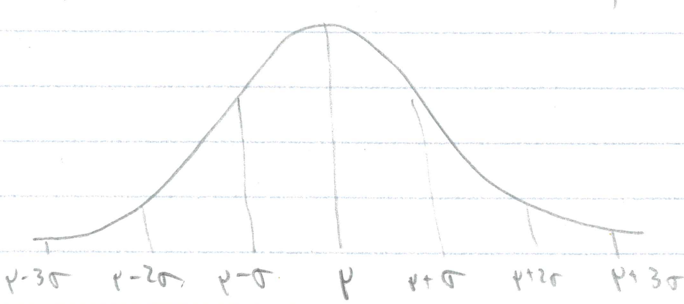
- Most common continuous distribution: Normal distribution

- $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\dfrac{1}{2\sigma^2}(x-\mu)^2\right]$ if $X \sim N(\mu, \sigma^2)$

↑ density

- Function of σ^2 (variance) and μ (mean)

- Also, If $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$, $Var(X) = \sigma^2$



- $\boxed{\text{Different Means and Different variances figure}}$

- The <u>standard</u> <u>normal</u> distribution is $N(0,1)$

- <u>Properties:</u>

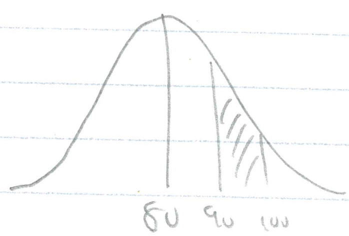
1.) 68 - 95 - 99.7 rule
   68% of area $\pm 1\sigma$
   95% of area $\pm 2\sigma$
   99.7% of area $\pm 3\sigma$
2.) Symmetric, $f(\mu - x) = f(\mu + x)$
3.) I median $= \mu$
4.) If $X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma} \Rightarrow Z \sim N(0,1)$
independent 5.) If $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, $Z = X + Y$
   then $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

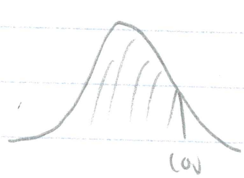
- We denote PDF of standard normal by $\phi(x)$
- The CDF is $\Phi(x) = P_r(X \leq x)$

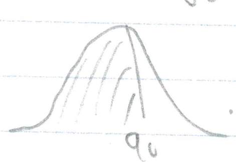
- Blood Pressure is $N(80, 144)$
  mild hypertensive is $90 \leq DBB \leq 100$
  Units are in mmHg
  What is P_r(mild hypertensive)?

Solution

$P_r(90 \leq X \leq 100) =$



$=$ 

$-$ 

$= P_r(X \leq 100) - P_r(X \leq 90)$

$=$ pnorm(100, mean = 80, sd = $\sqrt{144}$) $-$ pnorm(90, mean = 80, SD = $\sqrt{144}$)

$= 0.1545$

in inches

- **Exercise:** tree diameter $\sim N(8, 2^2)$
What is prob. tree has diameter > 12 in?

Solution: $1 - \text{pnorm}(12, 8, 2) = 0.02275$

- If $X_1, \ldots, X_n$ are random variables and

$L = \sum_{i=1}^{n} c_i X_i$ for c_i constants (Not RV's)

Then $E[L] = \sum_{i=1}^{n} c_i E(X_i)$

If the X_i's are independent then

$Var(L) = \sum_{i=1}^{n} c_i^2 Var(X_i)$

If the X_i's are also normal, then

$L \sim N(E(L), Var(L))$

- Ex.) X = serum creatinine level for Caucasian individual
Y = serum creatinine level for Black individual
$X \sim N(1.3, 0.25)$
$Y \sim N(1.5, 0.25)$
What is distribution of Average level for one Caucasian and one black individual chosen at random?

Solution
$Z = \frac{1}{2}X + \frac{1}{2}Y$, $Z \sim N(1.4, 0.125)$

$E(Z) = \frac{1}{2}1.3 + \frac{1}{2}1.5 = 1.4$

$Var(Z) = \frac{1}{4}0.25 + \frac{1}{4}0.25 = 0.125$

- Normal Approximation to Binomial

rule of thumb

If $X \sim Bin(n, p)$ and $np(1-p) \geq 5$, then

$$X \approx N(np, np(1-p))$$

- Let $X \sim Bin(n,p)$, $Y \sim N(np, np(1-p))$

then $Pr(a \leq X \leq b) \approx Pr(a - \frac{1}{2} \leq X \leq b + \frac{1}{2})$

continuity correction

we will use this for 2-sample binomial tests.

- Why? Let $T_1, T_2, \ldots, T_n$ be n independent Bernoulli trials

$$T_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

Then $X = T_1 + T_2 + \cdots + T_n$
$= \sum_{i=1}^{n} T_i$

Central Limit theorem says normal for large n

- Normal Approximation to Poisson

$X \sim Pois(\mu)$

$Y \sim N(\mu, \mu)$

Then for $\mu \geq 10$ (rule of thumb)

$$Pr(a \leq X \leq b) \approx Pr(a - \tfrac{1}{2} \leq Y \leq b + \tfrac{1}{2})$$

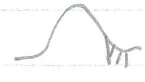
Exercise 5.12 – 5.13 of Rosner

- Of men 30-34 who have smoked, X = # years a man has smoked, Y = # of years smoked by women in age group

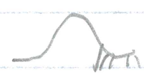
$X \sim N(12.8, 5.1^2)$, $Y \sim N(9.3, 3.2^2)$

Q.) What proportion of men have smoked for more than 20 years? women?

1 - pnorm(20, mean = 12.8, sd = 5.1)
= 0.07901



1 - pnorm(20, mean = 9.3, sd = 3.2)
= 0.004133



Exercise 5.126 – 5.130

Christmas Bird Count is a holiday tradition in a boring part of Massachusetts.

Year	# $= x_i$
2005	76
2006	47
2007	63
2008	53
2009	62
2010	69
2011	62

$\sum_{i=1}^{7} x_i = 432$ $\sum_{i=1}^{7} x_i^2 = 27{,}212$

1.) What is mean # of birds?

$\frac{1}{7} 432 = 61.71$

2.) What is sd #?

$\frac{1}{7}(27212) - \left(\frac{1}{7}432\right)^2 = 78.78$

$SD = \sqrt{78.78} = 8.876$

3.) Suppose # birds is normal w/ same mean and SD as parts 1 and 2. What is prob of at least 60 birds? Apply continuity correction.

1 − pnorm(59.5, mean = 61.71, sd = 8.876)

= 0.5983

4.) Find "Normal range", (L, U) integers such that
$L \le 15^{th}$ percentile,
$U \ge 85^{th}$ percentile

qnorm(c(0.15, 0.85), mean = 61.71, sd = 8.876)

= (52.51, 70.91)

52 to 80 is "normal range"

5.) What is probability $\# \ge U$ at least once during a 10-year period?

$Pr(X \ge U) \approx$ 1 − pnorm(79.5, mean = 61.71, sd = 8.876)

= 0.02252

$Y = \#$ years $\ge U \sim Bin(10, 0.02252)$

$Pr(Y \ge 1) = 1 - Pr(Y = 0) =$ 1 − dbinom(0, 10, 0.02252)

= 0.2037