Chapter 4 Class Notes

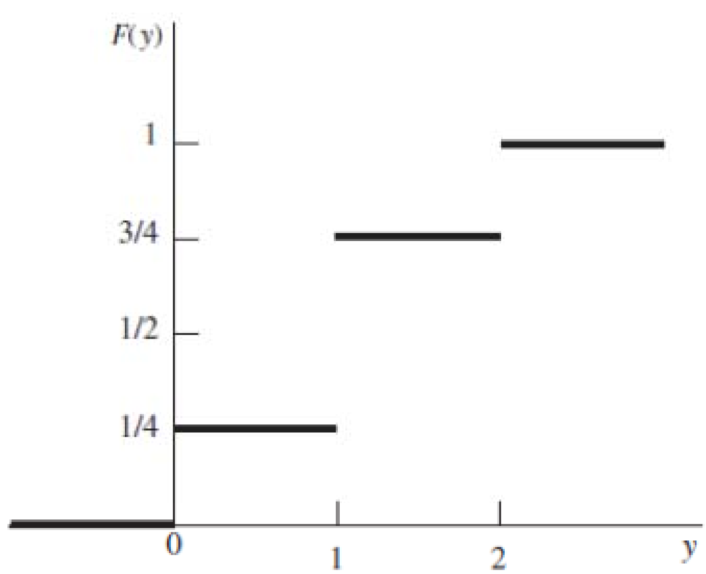
Sections 4.1 and 4.2

- **In Chapter 3, we looked at discrete distributions; here we study continuous ones. One example might be: suppose that day after day, we arrive at the Loyola EL stop at 8am, and the next train arrives with equal probability at any instance between 8:00 and 8:15.**
- **p.158, Def. 4.1: Let Y be any RV. Then F(y), the distribution function CDF of Y for -∞ < y < ∞, is**

$$\boldsymbol{F(y) = P(Y \leq y)}$$

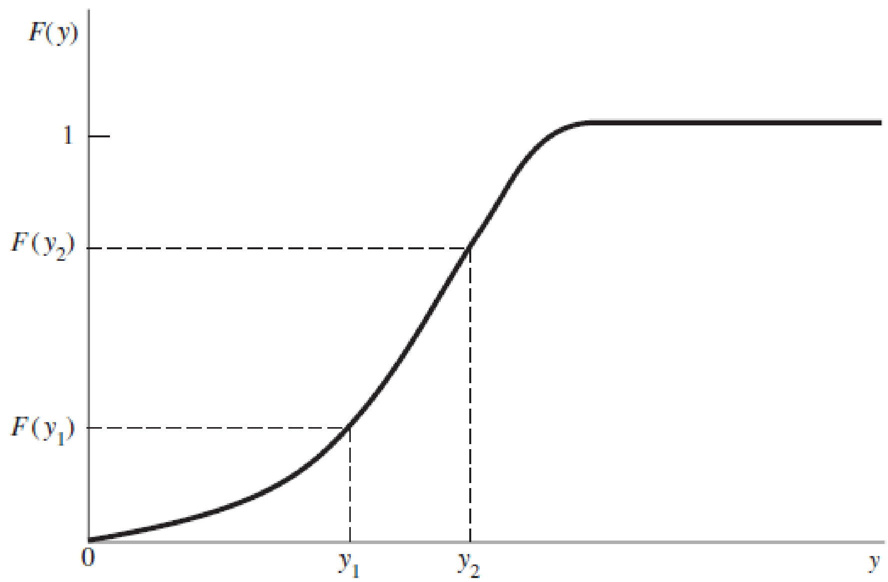
- **CDFs for discrete RVs are step functions (p.159), and for continuous RVs they are continuous (p.160)**
- **p.158, ex.4.1: Y ~ BIN(n=2,p=½), P(0)=P(2)=¼, P(1)= ½, so F(y) = 0 for y<0, F(y)=¼ for y in [0,1), F(y)=¾ for y in [1,2), and F(y)=1 for y≥2. Note the *right* continuity of F**

FIGURE 4.1 Binomial distribution function, $n = 2$, $p = 1/2$



- **The graph of a continuous CDF is as on p.160:**

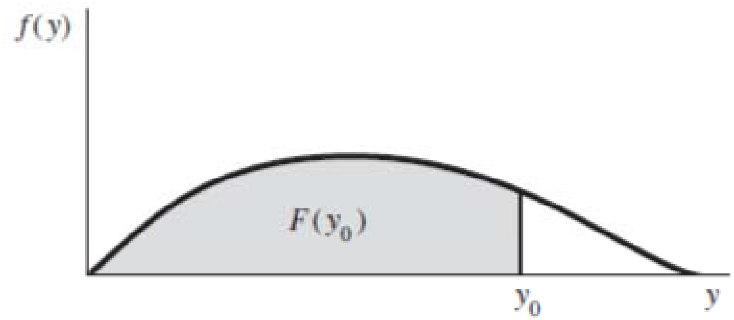
FIGURE 4.2 Distribution function for a continuous random variable



- **p. 160, Theorem 4.1: Properties of any CDF, F(x), are**
  - $\mathbf{F(-\infty) = \lim_{x \to -\infty} F(x) = 0}$
  - $\mathbf{F(\infty) = \lim_{x \to \infty} F(x) = 1}$
  - **Non-decreasing: For any real numbers y_1 and y_2 with $y_1 < y_2$, then $F(y_1) \leq F(y_2)$**
  - **Right continuous: $\forall y_0, F(y_0) = \lim_{y \to y_0^+} F(y)$**
- **p.160, Def. 4.2: The RV Y with CDF F(y) is said to be continuous if F(y) is continuous for all (real) y. For any continuous RV Y, note that $f(y_0) = P(Y = y_0) = 0$, so instead of probability *at a point*, in this case we think of probability *between points* and *areas under curves*.**
- **p.161, Def. 4.3: Let F(y) be the CDF for a continuous RV Y. Then f(y) given by f(y) = F'(y) is called the probability density function PDF for the RV Y. It follows also that $F(y) = \int_{-\infty}^{y} f(t)dt$**

- **This latter relationship is demonstrated in the following graph (from bottom of p.161):**

FIGURE 4.3 The distribution function



- **The PDF "is a theoretical model for the frequency distribution (histogram) of a population…" (p.161)**
- **p.162, Theorem 4.2: Let f(y) be the PDF for any continuous RV Y, then:**
  - **for all real y, f(y) ≥ 0**
  - $\int_{-\infty}^{\infty} f(y)dy = 1$
- **p.162, ex.4.2: Here, F(y) = 0 for y<0; F(y)=y for 0≤y<1; F(y) = 1 for y≥1, and the CDF and PDF are plotted in Figs. 4.4 and 4.5. The PDF f(y) is obtained by differentiation. This is an example of a CU (continuous uniform) RV over the interval (0,1) – see §4.4.**
- **p.163, ex.4.3: For this example, we start with the PDF, f(y) = $3y^2$ over 0<y<1 (and zero o/w) – graphed on p.163. We obtain the CDF by integration:**
  - **for y<0,** $F(y) = \int_{-\infty}^{y} 0dt = 0$
  - **for 0≤y<1,** $F(y) = \int_{-\infty}^{0} 0dt + \int_{0}^{y} 3t^2 dt = y^3$
  - **for y≥1,** $F(y) = \int_{-\infty}^{0} 0dt + \int_{0}^{1} 3t^2 dt + \int_{1}^{y} 0dt = 1$

- **p.164, Def. 4.4: Y is any RV and $p \in (0, 1)$, then the p^{th} quantile (or $100p^{th}$ percentile), ϕ_p, is the smallest value of Y such that $F(\phi_p) = P(Y \leq \phi_p) \geq p$.**
- **To illustrate, for the CDF from Ex.4.3, to find the median (50^{th} percentile), we can solve $\phi_{0.5}^3 = 0.5$ [ans. $= 0.5^{1/3} = 0.7937$] or use R:**

```
fn=function(y) y^3-0.5
uniroot(fn,c(0,1))
0.7937278
```

- **p.164, Theorem 4.3: Given a<b and continuous RV Y with PDF f(y), then the probability Y is in [a,b] is**

$$\boldsymbol{P(a \leq Y \leq b) = P(a < Y < b) = \int_a^b f(y)dy}$$

- **p.165, Exs.4.4/4.5: To make $f(y)=cy^2$ over [0,2] be a valid PDF, it must be the case that**

$$\boldsymbol{1 = \int_0^2 cy^2 dy = \frac{c}{3}y^3\Big]_0^2 = \frac{8}{3}c}$$

Thus, c = 3/8. Also,

$$\boldsymbol{P(1 \leq Y \leq 2) = \int_1^2 \frac{3}{8}y^2 dy = \frac{1}{8}y^3\Big]_1^2 = \frac{7}{8}}$$

Section 4.3

- **p.170, Def. 4.5: (Provided the absolute value version of the following integral exists) For the continuous RV Y, the expected value of Y is $E(Y) = \int_{-\infty}^{\infty} yf(y)dy$**

- **p.170, Theorem 4.4: (Provided the following integral exists for g(Y) a function of Y) The expected value of g(Y) is $E(g(Y)) = \int_{-\infty}^{\infty} g(y)f(y)dy$**
- **p.171, Theorem 4.5: As is true for the sum operator, so too the integral operator is a *linear operator*:**
$$E(af(Y) + bg(Y) + c) = aE(f(Y)) + bE(g(Y)) + c$$
- **As previously, for continuous RVs: $\sigma^2 = E(Y^2) - \mu^2$**
- **p.171, Ex.4.6: For Ex.4.2, $\mu = \int_0^2 \frac{3}{8}y^3 dy = \frac{3}{32}y^4\Big]_0^2 = 1.5$**

  Also, $E(Y^2) = \int_0^2 \frac{3}{8}y^4 dy = \frac{3}{40}y^5\Big]_0^2 = 2.4$, so $\sigma^2 = 2.4 - 1.5^2 = 0.15$, and $\sigma = 0.3873$.

Section 4.4

- **p.174, Defs. 4.6/4.7: For *parameters* $\theta_1 < \theta_2$, the RV Y has the continuous uniform CU distribution over the interval (θ_1,θ_2) if and only if the PDF of Y is $f(y) = \frac{1}{\theta_2-\theta_1}$ for $\theta_1 < y < \theta_2$ and f(y) = 0 o/w (otherwise).**
- **Application: When the number of calls coming into a switchboard over the time interval (0,t) has a Poisson distribution, given that one call came in during the interval (0,t), then we will see that the time the occurrence has the CU distribution over this interval. (We will prove this result later.)**

- **The CDF for the above CU distribution is:**

$$F(y) = \begin{cases} 0, & y < \theta_1 \\ \frac{y-\theta_1}{\theta_2-\theta_1} = -\frac{\theta_1}{\theta_2-\theta_1} + \frac{1}{\theta_2-\theta_1} y, & \theta_1 \le y < \theta_2 \\ 1, & y \ge \theta_2 \end{cases}$$

- **Note that the central portion of the CDF above is the line connecting the points (θ_1,0) and (θ_2,1)**
- **<u>p.175, Ex.4.7</u>: Given one [Poisson] occurrence in the interval (0,30) minutes, the probability the customer came during the last 5 minutes is**

$$P(25 \le Y \le 30) = F(30) - F(25) = 1 - \frac{25}{30} = \frac{1}{6}$$

In this case, $F(y) = \frac{y}{30}$ for 0 < y < 30.

- **P.176, <u>Theorem 4.6</u>: Y has the CU distribution on the interval (θ_1,θ_2). Then, $\mu = \frac{\theta_1+\theta_2}{2}$ and $\sigma^2 = \frac{(\theta_2-\theta_1)^2}{12}$**

<u>Section 4.5</u>

- **The Normal distribution is perhaps the most commonly-used dist. – and for good reason: the CLT!**
- **p.178, <u>Def. 4.8</u>: The RV Y has the Normal distribution with parameters μ and σ (μ is any real number and σ is any positive real number) IFF it has the PDF**

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \text{ for } -\infty < y < \infty$$

- **The Standard Normal PDF, $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$,is obtained by using the z-transformation: $Z = \frac{Y-\mu}{\sigma}$**
- **Since a closed-form solution is not possible, right-tail areas under the normal curve (obtained by numerical methods) are given in Table 4 (p.848 and on this page)**
- **Technically, we should write the Normal PDF as $f(y) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-a}{b}\right)^2}$ without specifying the roles of the two parameters, but we have the following:**
- **p.178, <u>Theorem 4.7</u>: For Y a Normal RV, E(Y) = μ and Var(Y) = σ^2**
- **Calling the above a (valid) PDF implies that**

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy = 1$$

  (this can be shown using a trick and polar coordinates)
- **<u>p.179, Ex.4.8</u>: For a Standard Normal RV, Z, we can obtain the listed probabilities (a. P(Z>2), b. P(-2≤Z≤2), and c. P(0≤Z≤1.73)) using both Table 4 and R:**

```
1-pnorm(2)
0.02275013
pnorm(2)-pnorm(-2)
0.9544997
pnorm(1.73)-pnorm(0)
0.4581849
```

- p.180, Ex.4.9: Y~N(μ=75,σ=10), then to use Table 4 to find P(80<Y<90), we first standardize: $\frac{80-75}{10} = 0.5$ and $\frac{90-75}{10} = 1.5$; then ans. = 0.3085 - 0.0668 = 0.2417
- In R: pnorm(90,75,10)-pnorm(80,75,10)
  0.2417303
- For Y a Normal RV with parameters μ and σ, the MGF is $m(t) = e^{\mu t + \sigma^2 t^2/2}$ $\forall t \in \mathbb{R}$. The proof of this result uses the "usual trick" –used often in this class.

Section 4.6

- p.185, Def. 4.9: The RV Y has the gamma distribution with (positive) parameters α and β IFF it has the PDF (for y≥0)

$$f(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}$$

- In this expression, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ is the so-called ***gamma function***; note: Γ(1)=1, $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$ for α>1, and for k an integer, Γ(k) = (k-1)!
- The typical graph of the gamma PDF is on p.185 (note the right skew), and additional plots are on p.186; α is the *shape parameter* and β is the *scale parameter*
- Showing that $\int_{-\infty}^{\infty} \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} dy = 1$ is simple by using the transformation x = y/b

- Note too that $\int_{-\infty}^{\infty} y^{\alpha-1}e^{-y/\beta}dy = \beta^{\alpha}\Gamma(\alpha)$
- Gamma probabilities and quantiles can be obtained in R by using pgamma(y_0,α,1/β) and qgamma(y_0,α,1/β)
- p.186, Theorem 4.8: For Y a gamma RV with parameters α and β, E(Y) = $\alpha\beta$ and Var(Y) = $\alpha\beta^2$
- p.187, Def. 4.10: For $\nu > 0$, the RV Y has the chi-square distribution with ν degrees of freedom (df) IFF Y has a gamma distribution with $\alpha = \nu/2$ and $\beta = 2$
- p.188, Theorem 4.9: Y a chi-square RV with ν degrees of freedom. Then μ = EY = ν and σ^2 = V(Y) = 2ν
- p.188, Def. 4.11: The RV Y has the exponential distribution with parameter β IFF its PDF for $y \geq 0$ is

$$f(y) = \frac{1}{\beta}e^{-y/\beta}$$

- p.188, Theorem 4.10: Y is an exponential RV with parameter β. Then μ = EY = β and σ^2 = V(Y) = β^2
- The CDF for an exponential (β) RV is F(y) = 1 - $e^{-y/\beta}$ and survival function P(Y>y) = $e^{-y/\beta}$
- p.188, Ex.4.10: Note that exponential RVs possess the ***memory-less property***:

$$P(Y > a + b | Y > a) = P(Y > b)$$

- For $t < 1/\beta$, the MGF for a gamma(α,β) RV is

$$m(t) = (1 - \beta t)^{-\alpha}$$

- For t < 1/2 the MGF for a chi-square(ν) RV is

$$m(t) = (1-2t)^{-\frac{\nu}{2}}$$

- For t < 1/β, the MGF for an exponential(β) RV is $m(t) = (1-\beta t)^{-1} = 1 + \beta t + \beta^2 t^2 + \beta^3 t^3 + \cdots$

Section 4.7

- p.194, Def. 4.12: The RV Y has the beta distribution with (positive) parameters α and β IFF it has the PDF (for 0 ≤ y ≤ 1)

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

- We define the beta function as

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- R can be used to find probabilities and quantiles using the commands pbeta(y_0,α,1/β) and qbeta(y_0,α,1/β)
- Since y lies in the interval from 0 to 1, the Beta distribution is often used for modelling proportions
- The RV X defined over the interval (a,b) can be transformed to this Beta interval using Y = (X-a)/(b-a); an example is p.173 ex. 4.32 where a=0 and b=4
- p.195, Theorem 4.11: Y is a beta RV with parameters α and β. Then $\mu = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- p.196, Ex.4.11: Y is the proportion sold during a week, modeled as Beta(α=4,β=2): f(y) = $20y^3(1-y)$; then

$P(Y > 0.9) = \int_{0.9}^{1} 20(y^3 - y^4)dy = y^4(5 - 4y)]_{0.9}^{1} = 1 - 0.9^4(1.4) = 0.08146$

Section 4.8: It is comforting to know that in practice, many statistical techniques (including ANOVA and regression) are *robust* to underlying distributional assumptions.

Section 4.9

- **p.202, Def. 4.13 just repeats previous (p.138) definitions of $\mu_k' = E(Y^k)$ (mean about the origin) and $\mu_k = E((Y - \mu)^k)$ (centered moment)**
- **p.202, Ex. 4.12: For Y~CU(θ_1=0,θ_2=θ), for k = 1, 2, …**

$$\mu_k' = E(Y^k) = \frac{1}{\theta}\int_0^{\theta} y^k dy = \frac{1}{(k+1)\theta} y^{k+1}\Big]_0^{\theta} = \frac{\theta^k}{k+1}$$

- **p.202, Def. 4.14 just repeats previous definition of the MGF $m(t) = E(e^{Yt})$ - note the constant 'b' must exist; similar results are given on top of p.203**
- **p.203, Ex. 4.13: As we saw previously, the MGF for a Gamma(α,β) RV is $m(t) = (1 - \beta t)^{-\alpha}$ for t < 1/β**
- **p.204, Ex. 4.14: By differentiation above, we obtain for a Gamma RV: $\mu_k' = \alpha(\alpha + 1)(\alpha + 2) \ldots (\alpha + k - 1)\beta^k$**
- **p.204, Ex. 4.15: Kinetic energy K, mass m and velocity v are related by K = ½mV^2 with V~Gamma(α=4,β=500). From the last exercise, E(V^2) = $\alpha(\alpha+1)\beta^2$ = 5,000,000, so E(K) = ½m(5,000,000) = 2,500,000m.**

- **p.205, Theorem 4.12: Y has PDF f(y), then the MGF for g(Y) is $E\left[e^{tg(y)}\right] = \int_{-\infty}^{\infty} e^{tg(y)} f(y)dy$**
- **Ex. 4.16: for Y~N(μ,σ), Y-μ has MGF $m(t) = e^{\sigma^2 t^2/2}$**

Section 4.10

- **p.205, Theorem 4.13 (restatement of Chebyshev's Thm.): Y has mean μ and variance σ^2, then for k > 0,**

$$P(|Y-\mu| < k\sigma) = P(\mu - k\sigma < Y < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

- **Note the proof of the above given on p.208**
- **p.208, Ex. 4.17: Y~Gamma(α=3.1,β=2); since μ = 6.2 and σ^2 = 12.4, y_0 = 22.5 is k = $\frac{22.5-6.2}{\sqrt{12.4}} = 4.63$ SD's above μ, so by Chebyshev's Theorem an upper bound for the (two-) tail probability is $1/4.63^2$ = 0.0467. The actual value from R is "2*(1-pgamma(22.5,3.1,0.5))" = 0.0023; this low p-value suggests that the new worker is slower than the previous workers.**

Section 4.11

- **Here we handle mixed (discrete and continuous) RVs and functions (such as expected values)**
- **p.210, Ex. 4.18: Y is daily petroleum sales (in 1000gal), and f(y) = $(3/8)y^2$, 0<y<2. The profit function is g(y) = 100y for 0<y<1 and g(y) = 140y for 1<y<2. Expected profit is then $206.25 by the calculation:**

$$E(g(Y)) = \int_0^1 100y \times \frac{3}{8} y^2 dy + \int_1^2 140y \times \frac{3}{8} y^2 dy$$

- **Now consider Y as a mixed distribution, mixing discrete and a continuous parts such as for the insurance example on p.211; then the CDF of Y is of the form $F(y) = cF_1(y) + (1-c)F_2(y)$. Here F_1 is the step portion, F_2 the continuous portion, and c is the accumulated probability of all the discrete points.**
- **p.211, Ex. 4.19: Y is electronic component life in 100 hours. P(Y=0) = ¼, and for y>0, Y~Exponential(1). Clearly, c = ¼, $F_1(y) = 0$ for y<0 and = 1 for y≥0. Also, $F_2(y) = 0$ for y<0 and $= 1-e^{-y}$ for y≥0. Thus, F(y) = 0 for y<0 and $F(y) = ¼ + ¾ (1-e^{-y}) = 1- ¾e^{-y}$ for y≥0. The survival function is $1-F(y) = ¾e^{-y}$, so $P(Y>10) = ¾e^{-10}$.**
- **p.212, Def. 4.15: Y is a mixed RV with CDF of the form $F(y) = cF_1(y) + (1-c)F_2(y)$. X_1 is a discrete RV with CDF $F_1(y)$ and X_2 is a continuous RV with CDF $F_2(y)$. g(Y) is a function of Y, then**

$$E[g(Y)] = cE(g(X_1)) + (1-c)E(g(X_2))$$

- **p.213, Ex. 4.20: To find the mean and variance for Ex. 4.19, note $E(X_1) = 0$, $E(X_2) = 1$, so E(Y) = ¼(0) + ¾(1) = ¾. Also, $E(X_1^2) = 0$, $E(X_2^2) = 2$, so $E(Y^2)$ = ¼(0) + ¾(2) = 3/2, and Var(Y) = 3/2 – $(¾)^2$ = 15/16. [Recall for X_2: $E(X_2) = \alpha\beta = 1$, $Var(X_2) = \alpha\beta^2 = 1$, so $E(X_2^2) = Var(X_2) + E(X_2)^2 = 1 + 1^2 = 2$.]**