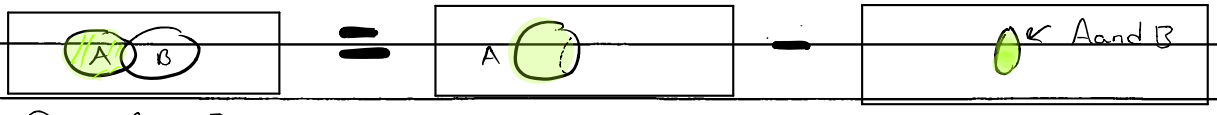
Rules Contin.

8) $Pr\{A \text{ and } B^c\} = Pr\{A\} - Pr\{A \text{ and } B\}$



9) $Pr\{A^c|B\} = 1 - Pr\{A|B\}$ or
$Pr\{B^c|A\} = 1 - Pr\{B|A\}$

BUT $Pr\{A|B^c\} \neq 1 - Pr\{A|B\}$ and $Pr\{B|A^c\} \neq 1 - Pr\{B|A\}$ } NOT TRUE

For example, if we know that out of all women, 25% wear glasses, we also know out of all women, 75% do not wear glasses ($Pr\{G^c|W\} = 1 - Pr\{G|W\}$)
However, we can not say that 75% of men wear glasses ($Pr\{G|W^c\}$).

Ex: Suppose the prob. of someone swimming regularly is 0.17, and the prob of someone swimming and getting an ear infection with a year is 0.09. Find the following:

a) The prob that someone swims and does not get an ear infection (within a year).

Let S = Swim, I = Infection.

$Pr\{S \text{ and } I^c\} = Pr\{S\} - Pr\{S \text{ and } I\} = 0.17 - 0.09 = 0.08$

b) If some one swims, the prob they get an infection.

$Pr\{I|S\} = Pr\{I \text{ and } S\} / Pr\{S\} = 0.09/0.17 = 0.529$

c) The prob. someone does not get an infection, if they swim.

$Pr\{I^c|S\} = 1 - Pr\{I|S\} = 1 - 0.529 = 0.471$

Definition: Two events A and B are independent if and only if either of these conditions holds true:

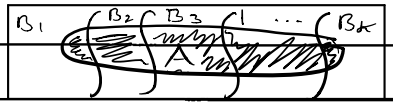
i) $Pr\{A \text{ and } B\} = Pr\{A\} Pr\{B\}$

ii) $Pr\{A|B\} = Pr\{A\}$

Independent means that the prob. of one event does not effect the prob. of another. The equations i) and ii) can only be used if you know two events are independent. Generally, we do not assume they are.

Last rule:

10) For any two events A and B,
$P_r\{A\} = P_r\{A \text{ and } B\} + P_r\{A \text{ and } B^c\}$
For any events A and $B_1, B_2, B_3, B_4, \ldots, B_k$,
$P_r\{A\} = P_r\{A \text{ and } B_1\} + P_r\{A \text{ and } B_2\} + \cdots + P_r\{A \text{ and } B_k\}$



We are just adding up all parts of B that overlap with A, to get A.
This rule is sometimes called the "rule of total prob".

Ex: The probability of a certain diseas is 0.05. If a person has the disease, the prob of testing positive is 0.99. If they do not have the disease, the prob of testing positive is 0.02.

a) Write down in probability notation the values you have been given.
Let + = positive, D = disease.
Given: $P_r\{D\} = 0.05$, $P_r\{+ \mid D\} = 0.99$, $P_r\{+ \mid D^c\} = 0.02$.
We can also get: $P_r\{D^c\} = 0.95$, $P_r\{- \mid D\} = 0.01$, $P_r\{- \mid D^c\} = 0.98$

b) Find the prob that someone tested positive.
$P_r\{+\} = P_r\{+ \text{ and } D\} + P_r\{+ \text{ and } D^c\}$ (using rule 10)
$= P_r\{+ \mid D\}P_r\{D\} + P_r\{+ \mid D^c\}P_r\{D^c\}$ (using rule 7 twice)
$= (0.99)(0.05) + (0.02)(0.95) = 0.0685$
* Why is this prob. larger than $P_r\{D\}$?

c) What is the prob that someone who tested + actually has the disease?

$$P_r\{D|+\} = \frac{P_r\{+ \text{ and } D\}}{P_r\{+\}} = \frac{P_r\{+|D\}P_r\{D\}}{P_r\{+\}} = \frac{0.99(0.05)}{0.0685} = 0.7226$$

I.e., if you have tested positive, there is a 72.26% chance you have the disease.

d) What is the probability that someone tests positive, or has the disease, or both?

$P_r\{+ \text{ or } D\} = P_r\{+\} + P_r\{D\} - P_r\{+ \text{ and } D\} = 0.05 + 0.0685 - 0.99(0.05)$

$= 0.069$

e) Are testing positive and having the disease independent?

To check this, we would see if

i) $P_r\{+ \text{ and } D\} = P_r\{+\}P_r\{D\}$ or ii) $P_r\{+|D\} = P_r\{+\}$

(You can check any other combinations as well

For i): $P_r\{+\}P_r\{D\} = 0.0685(0.05) = 0.003425$

$P_r\{+ \text{ and } D\} = 0.99(0.05) = 0.0495$

These are not equal, so they are not independent (They are dependent).

For ii) $P_r\{+|D\} = 0.99$, $P_r\{+\} = 0.0685$, which are not equal.

(note; you only needed to check one of these).

Discrete Random Variables

Recall: A discrete r.v. is numeric with natural gaps.

Fact: All the probability rules we learned also apply to all r.v's.

Ex: Let Y = # on a die. Then, Y can equal 1, 2, 3, 4, 5, 6, so is a discrete r.v.

Notice all values of Y are mutually exclusive. This is true for all r.v's.

We can use our probability rules to calculate:

a) $P_r\{Y > 5\}$ (we roll a number > 5) $= P_r\{Y = 6\}$ (because of the gap) $= 1/6$ (if it is a fair die)

"A^c" $1 - P_r\{A\}$

b) $P_r\{Y \geq 1\} = 1 - P_r\{Y < 1\} = 1 - P_r\{Y = 0\} = 1 - 1/6 = 5/6$

c) $P_r\{Y \leq 1 \text{ or } Y \geq 5\} = P_r\{Y \leq 1\} + P_r\{Y \geq 5\} + 0$ (mutually exclusive)
$= P_r\{Y = 1\} + P_r\{Y = 5\} + P_r\{Y = 6\} = 3/6.$

In addition, for any discrete r.v we have,

Definition: The mean of a discrete r.v Y is (or expected value)

$\mu_Y = \sum_{i=1}^{k} y_i P_r\{Y = y_i\}$, Y takes on values $y_1, y_2, \ldots, y_k$

Definition: The variance of a discrete r.v Y is

$\sigma_Y^2 = \sum_{i=1}^{k} (y_i - \mu_Y)^2 P_r\{Y = y_i\} = [\Sigma y_i^2 P_r\{Y = y\}] - (\mu_Y)^2$

For the dice example, we have

$\mu_Y = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 3.5$

$\sigma_Y^2 = [1^2(1/6) + 2^2(1/6) + 3^2(1/6) + 4^2/1/6) + 5^2(1/6) + 6^2(1/6)] - 3.5^2$
$= 15.166 - 12.25 = 2.916$

Ex: In a year of observing patients, the distribution of time to "full recovery" after knee replacement was:

# months	6	12	18	24
Rel. Freq	0.20	0.30	0.35	0.15

You may assume this is population data.

Let Y = # months to full recovery.

a) Find the average of Y.

$\mu_Y = \sum_{y_i} y_i \, Pr\{Y = y_i\} = 6(0.20) + 12(0.30) + 18(0.35) + 24(0.15)$

$= 14.7$ months

b) Find the variance of Y.

$\sigma_Y^2 = \sum_{y_i} (y - \mu_Y)^2 \, Pr\{Y = y_i\} = (16-14.7)^2(0.20) + (12-14.7)^2(0.30) +$

$(18-14.7)^2(0.35) + (24-14.7)^2(0.15) = 34.11$

c) Interpret the standard deviation in terms of the problem.

$\sigma_Y = \sqrt{34.11} = 5.84$, a typical distance/deviation from the mean is 5.84 months.

* Linear Combinations of Random Variables

If X is a r.v. with mean μ_X, std dev σ_X, then

i) If $Y = X + a$ (we add a constant "a" to all X values)

$\mu_Y = \mu_X + a$, $\sigma_Y = \sigma_X$

ii) If $Y = bX$ (we multiply every X value by a constant "b")

$\mu_Y = b\mu_X$, $\sigma_Y^2 = b^2\sigma_X^2$, $\sigma_Y = |b|\sigma_X$

These give the last result,

iii) If $Y = a + bX$ (a linear combination)

$\mu_Y = a + b\mu_X$, $\sigma_Y^2 = b^2\sigma_X^2$, $\sigma_Y = |b|\sigma_X$

Ex: If X has $\mu_X = 4$, $\sigma_X = 2$, and $Y = 2 - 4X$

Find μ_Y and σ_Y^2.

$a = 2$, $b = -4$, $\mu_Y = 2 + (-4)\mu_X = -12$

$\sigma_Y^2 = b^2\sigma_X^2 = (-4)^2(2^2) = 64$

* These properties apply to <u>any</u> r.v.

Binomial Random Variables

A very frequently used r.v. is one with the following properties

1) An experiment is conducted with exactly two possible outcomes, one of which is a "success", one a "failure".
2) The outcomes of all experiments/trials are independent.
3) The probability of a "success" for a single trial is denoted p, and p does not change.
4) The variable of interest is Y = # of successes in n trials

Then, Y is a binomial r.v, which can take on values 0, 1, 2, 3, ..., n.
The parameters of a binomial distribution are n = # of trials, p = Pr{"success"}.

Ex: A coin flipped 5 times, where a "success" is a head.
n = 5, p = 0.50

Ex: The number of children who are over 65 inches tall, out of 4. n = 4, p = Pr{over 65 inches}

Ex: A couple has 3 children, and each child has a 15% chance of having blue eyes.

a) Identify the distribution, and its parameters, if we are interested in the # of children with blue eyes.
Y = # of children with blue eyes out of 3, n = 3, p = 0.15. This is a binomial r.v., Y can = 0, 1, 2, 3

b) Let: B_i = event child i has blue eyes. Find the probability that no children have blue eyes.

We want:

NOTE: $P_r\{B_i^c\} = 1 - P_r\{B_i\}$

$P_r\{B_1^c \text{ and } B_2^c \text{ and } B_3^c\} \overset{\text{by independence}}{=} P_r\{B_1^c\} P_r\{B_2^c\} P_r\{B_3^c\}$,

$= (1-.15)(1-.15)(1-.15) = 0.614$

c) Find the probability that exactly two children have blue eyes.

$P_r\{Y=2\} = P_r\{(B_1 \text{ and } B_2 \text{ and } B_3^c) \text{ or } (B_1 \text{ and } B_2^c \text{ and } B_3) \text{ or } (B_1^c \text{ and } B_2 \text{ and } B_3)\}$

(Mutually Ex.) $= P_r\{B_1 \text{ and } B_2 \text{ and } B_3^c\} + P_r\{B_1 \text{ and } B_2^c \text{ and } B_3\} + P_r\{B_1^c \text{ and } B_2 \text{ and } B_3\}$

$= P_r\{B_1\} P_r\{B_2\} P_r\{B_3^c\} + P_r\{B_1\} P_r\{B_2^c\} P_r\{B_3\} + P_r\{B_1^c\} P_r\{B_2\} P_r\{B_3\}$

$= (0.15)(0.15)(0.85) + (0.15)(0.85)(0.15) + (0.85)(0.15)(0.15) = 3(0.15^2)(0.85)$

$= 0.0574$

I.e,

$P_r\{Y=2\} = 3(0.15^2)(0.85)$

3 = # of ways we can get 2 blue eyed children out of 3

$(0.15)^2$ = prob of getting 2 blue eyed children

(0.85) = prob of getting 1 non-blue eyed child

Notice we used that values of Y are mutually exclusive, while the "trials" (children) are independent.

If n were larger, it would be tedious to calculate the above. However, a binomial r.v. has simplified equations for calculating $P_r\{Y=j\}$, where j is some number:

* Properties of binomial random variables

Let Y be a binomial r.v with parameters n, p. Then the following hold true:

i) $P_r\{Y=j\} = \binom{n}{j} p^j (1-p)^{n-j}$, for $j = 0, 1, 2, \ldots, n$

= prob of j successes in n trials, where

$\binom{n}{j} = \frac{n!}{j!(n-j)!}$, where $n! = n(n-1)(n-2)\cdots(3)(2)(1)$

Other notation for $\binom{n}{j}$ is ${}_nC_j$. NOTE: $0! = 1$.

ii) μ_Y = expected / average # of successes in n trials

$= np$

iii) σ_Y^2 = variance of the # of successes in n trials

$= np(1-p)$

Ex of n!: $4! = 4(3)(2)(1) = 24$, $\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(2 \cdot 1)} = 3 \cdot 2 \cdot 1 = 6$

Ex: A plant has a 20% chance of having a red flower bloom on any bud. If a plant has 6 buds, find the following:

a) The prob. of exactly one red flower.

Y is binomial, $n = 6$, $p = 0.20$.

$P_r\{Y = 1\} = \binom{6}{1}(.20)^1(1-.20)^5 = \frac{6!}{1!(6-1)!}(.20)(.80)^5 = \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{(1)5\cdot4\cdot3\cdot2\cdot1}(.0655)$

$= 0.393$

b) The prob of at least two red flowers.

$P_r\{Y \geq 2\} = 1 - P_r\{Y < 2\} = 1 - P_r\{Y \leq 1\} = 1 - (P_r\{Y=1\} + P_r\{Y=0\})$

$= 1 - (0.393) - \binom{6}{0}.20^0(1-.20)^6 = 1 - 0.393 - .80^6$

$= 1 - 0.393 - 0.262 = 0.345$

c) The mean and variance for the # of red flowers out of 6

$\mu_Y = np = 6(.20) = 1.2$, $\sigma^2_Y = np(1-p) = 6(.20)(.80) = .96$

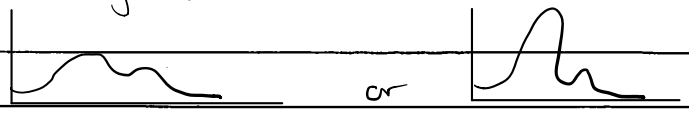
Note: $\binom{n}{0} = 1$, $\binom{n}{1} = n$, $\binom{n}{n-1} = n$, for any n.

Do not round μ_Y, σ^2_Y.

Probabilities for Continuous Random Variables

With continuous data, we cannot use the same approach as we did with discrete, since there are an INFINITE number of values.

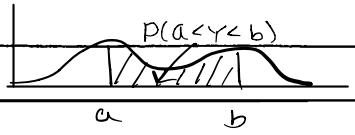
We instead find probabilities as areas under density curves. You can think of a density curve as a smoothed histogram:

 or

The curve is the function that creates the graph.

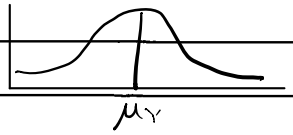
Density curves have the following properties:

1) The total area under the curve MUST be 1
2) Probabilities are areas under curves (found via calculus)
3) The area under the curve between points "a" and "b" gives the prob. the r.v. takes on values between "a" and "b".



Normal Random Variables

The most widely used continuous r.v. in STATISTICS is the Normal / Gaussian r.v. It is perfectly symmetric about its mean, and is "bell shaped":



Some facts about a normal r.v.:

1) $\bar{y} = \tilde{y}$ (The mean equals the median)
2) $Pr\{Y > \mu + a\} = Pr\{Y < \mu - a\}$  (Symmetric)

3) If Y is distributed normal with mean μ_Y, std dev σ_Y, we denote it as $Y \sim N(\mu_Y, \sigma_Y)$

Standardizing data

For any random variable Y with mean μ_Y, std. dev. σ_Y, if we use the following linear transformation:

$Y^* = (Y - \mu_Y)/\sigma_Y$, then $\mu_{Y^*} = 0$, $\sigma_{Y^*} = 1$.

To show this:

$Y^* = Y/\sigma_Y - \mu_Y/\sigma_Y$, $a = \mu_Y/\sigma_Y$, $b = 1/\sigma_Y$, so $\mu_{Y^*} = \frac{\mu_Y}{\sigma_Y} - \frac{\mu_Y}{\sigma_Y} = 0$

$\sigma^2_{Y^*} = b^2\sigma^2_Y = (1/\sigma_Y)^2\sigma^2_Y = 1 \Rightarrow \sigma_{Y^*} = 1$

Standard Normal Random Variable

When $Y \sim N(\mu_Y, \sigma_Y)$, the transformation $(Y-\mu_Y)/\sigma_Y$ is also normal, with mean 0, std. dev 1. This is called a standard normal r.v., and typically denoted with Z.

I.e. If $Y \sim N(\mu_Y, \sigma_Y)$, and $Z = (Y-\mu_Y)/\sigma_Y$ then $Z \sim N(0,1)$

The "Z-table" in your book has calculated areas under curves for $Z \sim N(0,1)$. The table gives $Pr\{Z < Z^*\}$, $Z^* = -3.49, -3.48, \ldots, 3.48, 3.49$

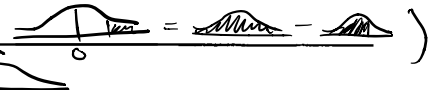
$Z = (Y-\mu_Y)/\sigma_Y$ can be calculated on a single data point, at which point we call it a z-score.

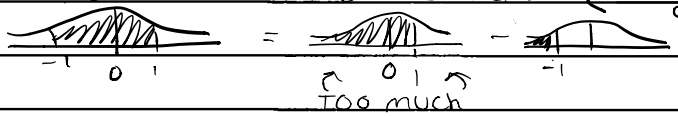
z-score $= (y-\mu_Y)/\sigma_Y$ = The # of std dev.'s y is from μ_Y.

The table in the back gives the second decimal place as columns, the first two digits as rows, and probabilities in the middle.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	$Pr\{Z<0.00\}$				$Pr\{Z<0.04\}$					
0.1										
⋮										
1.0	0.8413									
1.1					$Pr\{Z<1.15\}$					
⋮										
3.4										$Pr\{Z<3.49\}$

Ex: $Pr\{Z<1\}$ can be found directly in row 1.0, column 0.00 = 0.8413

$Pr\{Z>1\} = 1 - Pr\{Z<1\} = 1 - 0.8413 = 0.1587$ ()

$Pr\{-1<Z<1\}$ = 

$= Pr\{Z<1\} - Pr\{Z<-1\} = 0.8413 - 0.1587 = 0.6826$

(Notice $Pr\{Z>1\} = Pr\{Z<-1\}$ by symmetry)

Percentiles for a Z, and $Z_{\alpha/2}$

The k^{th} percentile for a Z can be found by looking up $k/100$ (or the closest value to it) in the middle of the Z table, then seeing what Z-score that is by looking at what row and column it was in.

Ex: Find the ~70^{th} percentile for a Z.

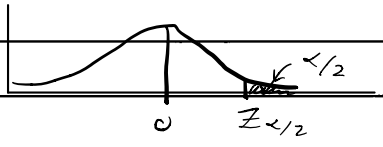
0.6985 is in column .02, row 0.50, so $P_r\{Z < .52\} = 0.6985$

0.7019 is in column .03, row 0.50, so $P_r\{Z < .53\} = 0.7019$

Since .6985 is closer to 0.70, the approximate 70^{th} percentile for Z is $Z^{(70)} = .52$.

Note: If two values in the middle are equi-distance from $k/100$, average the Z-scores.

$Z_{\alpha/2}$ is defined to be the value of Z with $\alpha/2$ in the upper tail;



So for ex, $Z_{.10}$ has 0.10 in the upper tail, 0.90 in the lower (it is the 90^{th} percentile).

We will use this notation later.

* Using the Z-table for all $Y \sim N(\mu_Y, \sigma_Y)$

When you have $Y \sim N(\mu_Y, \sigma_Y)$, and want any probability of Y, you can transform Y into $Z \sim N(0,1)$ then use the Z-table:

i) $P_r\{Y < a\} = P_r\{(Y-\mu_Y)/\sigma_Y < (a-\mu_Y)/\sigma_Y\} = P_r\{Z < \frac{(a-\mu_Y)}{\sigma_Y}\}$

ii) $P_r\{Y > a\} = 1 - P_r\{Y < a\} = 1 - P_r\{\frac{(Y-\mu_Y)}{\sigma_Y} < \frac{(a-\mu_Y)}{\sigma_Y}\} = 1 - P_r\{Z < \frac{(a-\mu_Y)}{\sigma_Y}\}$

iii) $P_r\{a < Y < b\} = P_r\{Y < b\} - P_r\{Y < a\} = P_r\{Z < \frac{(b-\mu_Y)}{\sigma_Y}\} - P_r\{Z < \frac{(a-\mu_Y)}{\sigma_Y}\}$

iv) To find a percentile for Y, first find the percentile for Z, then solve for Y:

$$Z^{(k)} = \frac{Y^{(k)} - \mu_Y}{\sigma_Y} \Rightarrow Y^{(k)} = \mu_Y + Z^{(k)}(\sigma_Y)$$

Ex: The distribution for the score on an exam is normal, with mean 60, std. dev 10. Find the following:

a) The prob. that a student scores over 80.

Y = exam score, $Y \sim N(60, \sigma_Y = 10)$

$$P_r\{Y > 80\} = P_r\left\{\frac{Y-\mu_Y}{\sigma_Y} > \frac{(80-60)}{10}\right\} = P_r\{Z > 2\} = 1 - P_r\{Z < 2\}$$
$$= 1 - 0.9772 = 0.0228$$

b) The prob of scoring within two std dev's of the mean.

$$P_r\{60 - 2(10) < Y < 60 + 2(10)\} = P_r\{40 < Y < 80\} = P_r\{Y < 80\} - P_r\{Y < 40\}$$
$$= P_r\left\{Z < \frac{80-60}{10}\right\} - P_r\left\{Z < \frac{40-60}{10}\right\} = P_r\{Z < 2\} - P_r\{Z < -2\}$$
$$= 0.9772 - 0.0228 = 0.9544$$

c) The 90^{Th} percentile of exam score.

$Z^{(90)} = 1.28$ (.90 is in row 1.2, column 0.08)

So, $Y^{(90)} = \mu_Y + Z^{(90)}(\sigma_Y) = 60 + 1.28(10) = 72.8$

I.e, 10% of students scored over 72.8, 90% below.