Introduction to Probability and Statistics

Professor Mark Brown
Statistics Department
Columbia University
Notes by Yiqiao Yin in LaTeX

December 13, 2016

Abstract

This is the lecture notes from Prof. Mark Brown Introduction to Probability and Statistics, an upper level course offered in Fall 2016. A calculus-based tour of the fundamentals of probability theory and statistical inference. Probability models, random variables, useful distributions, conditioning, expectations, law of large numbers, central limit theorem, point and confidence interval estimation, hypothesis tests, linear regression.

This note is dedicated to Professor Mark Brown.

Contents

1 Descriptive Statistics

Go back to Table of Contents. Please click TOC

In 1627, Grand Duke of Tuscany consulted with Galileo the following problem.

Roll 3 dices. The common wisdom among gamblers is that the sum is more likely to be 10 than 9. Is that correct? Galileo (1564-1642) gave the following answer. The outcome can be viewed as (i, j, k) with $i = 1, ..., 6$, $j = 1, ..., 6$, and $k = 1, ..., 6$. All $6 \times 6 \times 6 = 216$ outcomes (sample points) are equally likely to happen. Let $A = \{A\ collection\ of\ sample\ points\}$, which is now called an event, we have

$$P(A) = \frac{Number\ of\ Sample\ Points}{216}$$

We can denote events with A_i and let i be integers. That is, for $A_1 = \{sum = 9\}$, we have 6 sample points for $621, 134, 234$, 3 sample points for $144, 225$, and 1 sample point for 333. Thus, $P(A_1) = \frac{25}{216}$. For $A_2 = \{sum = 10\}$, we have 6 sample points for $631, 541, 532$, and 3 sample points for $622, 433, 442$. Thus, $P(A_2) = \frac{27}{216}$. Therefore, the final answer is very clearly, $P(A_2) - P(A_1) = \frac{1}{108}$.

Another story happened with Pascal (1623 - 1662) and Fermat (1601 - 1665) in 1654 about flipping coins. The first player to win 3 games gets 32 pistoles. Also they assumed that it was an even game (with fair coins). There was also quitting fees. That is, a player trial (0,1) or (0,2) or (1,2). The question is, "what is a fair quitting fee?"

The solution is as the following.

Let $p = Pr(player\ wins\ given\ present\ position)$. Player would expect change of fortune when the game ends. That is $32p + (-32)(1 - p) = 64p - 32$, that is, the fair quitting fee would be $32 - 64p$. For the event (0,1), we have $p = Pr\{WWW, WLWW, LWWW, WWLW\} = Pr(WWW(= \frac{1}{8}), WLWW(= \frac{3}{16}), LWWW(= \frac{3}{16}), WWLW(= \frac{3}{16})\} = \frac{5}{16}$, then we have $p - q = \frac{5}{16} - \frac{11}{16} = -\frac{6}{16}$. Then we have $32(p - 2) = -12$. Thus the fair quitting fee in this case is 12 pistoles. For the event (1,2), $p = Pr(WW) = \frac{1}{4}$. Then $p - q = \frac{1}{4} - \frac{3}{4} = -\frac{1}{2}$. Then $32(p - q) = -16$. The fair quitting fee would be 16 pistoles. For the event (0,2), $p = Pr(WWW) = \frac{1}{8}$. We have $p - q = -\frac{6}{8}$. Then $32(p - q) = -24$. The fair quitting fee would be 24 pistoles.

1.1 General Problem

Go back to Table of Contents. Please click TOC

We can develop the problem above into a general problem. We can play series of independent games. Let us say the probability α of winning in each trial (win \$1 or lose \$1 with probability α and $1 - \alpha$). Find the probability that you win m trials before you lose n trials. In quitting problem, $\alpha = \frac{1}{2}$. In (1,2) case, $m = 2$, and $n = 1$.

Here is a solution by Pascal. Imagine that the game continues beyond the decision outcome. We can look at the next $m + n - 1$ trials. We would either have the case of winning at least m trials or the case of winning

at most $m-1$ trials. For the first case, winning at least m trials, the number of wins $\geq m$ and the number of losses $\leq n-1$, so the player can win m before losing n trials. For the alternative case, the player wins at most $m-1$ trials, we have the number of wins $\leq m-1$, and the number of loses $\geq n$. Then the player does not win m before losing n. To sum up, the probabilities for the two cases are equivalent to each other, $Pr(win\ m\ before\ losing\ n) = Pr(\geq m\ wins\ in\ m+n-1\ trials)$.

We can also discuss this situation with Binomial distribution as the following

$$Pr(exactly\ k\ wins\ in\ m+n-1\ trials)$$
$$= \sum_{k=m}^{m+n-1} \binom{m+n-1}{k} \alpha^k (1-\alpha)^{m+n-1-k}$$

Let us look at a quitting problem. A trial can be from 0 to 1. Let $\alpha = \frac{1}{2}$. We have $P(A\ wins\ match) = Pr(3\ wins\ before\ 2\ losses)$. That is, $m=3$, $n=2$, $m+n-1=4$, and $\alpha = \frac{1}{2}$. Thus,

$$\sum_{k=3}^{4} \binom{4}{k} (\frac{1}{2}^k)(\frac{1}{2})^{4-k} = \frac{1}{16}(4+1) = \frac{5}{16}.$$

Example 1.1. *Here is another example. Team A plays team B two games to none in a best 4 of 7 series. Find the probability that A wins series. Using the model above with* $\alpha = \frac{1}{2}$*, and we have* $m=4$*,* $n=2$*,* $m+n-1 = 5$*, and we have*

$$\sum_{k=4}^{5} \binom{5}{k} (\frac{1}{2}^k)(\frac{1}{2})^{5-k} = \frac{1}{32}(5+1) = \frac{3}{16}.$$

Note that the assumptions here can be questionable.

Roll a pair of dice 24 times. Win 1 unit if roll at least one (6,6); otherwise lose 1 unit. Should you bet on win or lose? For this problem, the Chevelier de Mere consulted with Pascal. The Chevelier thought that it was more likely to lose.

The solution given is as the following. $Pr(lose) = (\frac{35}{36})^{24} \approx 0.5086$, which was a difficult calculation at the time. Pascal remarked to Fermat that the Chevalier de Mere was such an assiduous gambler that he could distinguish empirically between 0.4914 and 0.5086.

For Christian Hoygens (1629 - 1695), we had the discover of moons of Saturn and he was also the inventor of pendulum clock.

Here is another interesting problem. There are 12 balls with 4 white and 8 black. A goes first, B second, and C third. Sample balls with replacement. Game ends as soon as a player chooses a white ball. Find appropriate stakes to make the game fair.

Isaac Newton (1643 - 1727) was another famous mathematicians, physicists, and a lot of things that we know of.

Samuel Pepys (1633 - 1703), an English member of the Parliament and a nobleman, consulted Newton on a dice problem. At least one 6 in 6 rolls of a fair die or at least two 6's in 12 rolls or at least three 6's in 18 rolls. Which is the biggest?

The solution goes

$$Pr(\geq 1 \ six \ in \ 6 \ rolls) = 1 - (\frac{5}{6})^6 \approx 0.665$$

$$Pr(\geq 2 \ sixes \ in \ 12 \ rolls) = 1 - \left((\frac{5}{6})^{12} + 12(\frac{5}{6})^{11}\frac{1}{6}\right) \approx 0.6187$$

$$Pr(\geq 3 \ sixes \ in \ 18 \ rolls)$$

$$= 1 - \left((\frac{5}{6})^{18} + 18(\frac{5}{6})^{17}(\frac{1}{6}) + \frac{18 \times 17}{2}(\frac{5}{6})^{16}(\frac{1}{6})^2\right) \approx 0.5973$$

John Arbuthnot (1667 - 1735), a mathematician, physicist, physician, scholar, satirist and author invented the term probability. He translated Hoygens book on probability, which was the first work on probability in the English language. "When mathematical reasoning can be had, it is as great a folly to make use of any other as to grope far a thing in the dark when you have a candle standing by you."

Arbuthnot wrote about the superstition that 13 was an unlucky number. He discussed the death rate was 1 in 26 per year. The story goes as the following. If 13 people are in a room, their combined death rate is $\frac{1}{2}$. In going from 12 to 13, the rate goes from below $\frac{1}{2}$ to $\frac{1}{2}$. Let us say there are k people in the room. Then $Pr(at \ least \ 1 \ will \ die \ in \ the \ next \ year) = 1 - (\frac{25}{26})^k = 1 - (\frac{25}{26})^k$, and in fact $1 - (\frac{25}{26})^k > \frac{1}{2}$. If $1 - (\frac{25}{26})^k < \frac{1}{2} \Leftrightarrow k \log \frac{26}{25} > \log 2 \Leftrightarrow k > (\log 2)/(\log \frac{26}{25}) \approx 17.67$. If $k = 17$, $Pr(at \ least \ 1 \ death) \approx 0.4866$. If $k = 18$, $Pr(at \ least \ 1 \ death) \approx 0.5061$.

We also have the birthday problem. For a group of $r \geq 2$ people. We assume all 365 (not Feb. 29) birthdays are equally likely, and that the birthdays of different people in the group are independent of one another. Find the smallest value of r such that $P(r) = Pr(\text{at least 2 people have the same birthday}) > \frac{1}{2}$.

The solution is the following. $q(r) = 1 - p(r) = Pr(a)$, and r birthdays are different. We want the smallest r such that $q(r) < \frac{1}{2}$. Then we have

$$q(r) = \frac{364}{365} \cdots \frac{365 - (r-1)}{365} = \prod_1^{r-1}\left(\frac{365-k}{365}\right) = \prod_j^{p-1}(1 - \frac{j}{365}).$$

Then we have $\Rightarrow q(22) = 0.5243$, $\Rightarrow q(23) = 0.4977$, $\Rightarrow p(23) = 0.5073$. Note that $q(r) = \prod_i^{p-1}(1 - \frac{k}{365}) = exp(\sum_i^{r-1} \log(1 - \frac{k}{365})) \approx exp(\sum_1^{r-1} \frac{k}{365}) = exp(-\frac{r(r-1)}{730})$. Then we want $exp(\frac{r(r-1)}{730}) < \frac{1}{2} \Leftrightarrow \frac{r(r-1)}{730} > \log 2$

$$\Rightarrow r^2 - r - 730 \log 2 > 0$$

$$\Rightarrow r > \frac{1 + \sqrt{1 + 2920 \log 2}}{2} = 22.9999$$

Hence, this suggests that $r \geq 23$ is the answer.

Example 1.2. *Another examples is to calculate the possibility of the event of full houses. We have* $13(12)(4)(6) = 3744$ *to be the total number of the times this event could happen. Then we have the probability,* $Pr(Full \ house) = \frac{3744}{2,598,960} \approx 0.00144 \approx= \frac{1}{694}$.

There is also a Gambler's Ruin Problem. Let us play a series of games. On each win of number 1 with probability p, lose of number of 1 with probability $q = 1 - p$. Start with r dollars and we play until either you reach 0 or n with $1 \leq r \leq n-1$, whichever comes first. The goal is to find $p(r) = Pr(0\ before\ n|start\ with\ r)$.

THe solution is as follows. We have $q(r) = 1-p(r)$, $q(0) = 0$, $q(n) = 1$. Then for $1 \leq j \leq n-1$, we have $q(j) = pq(j+1) + qq(j-1)$. Then $p(q(j+1) - q(j)) = q(q(j) - q(j-1))$. For $\triangle(j) = q(j+1) - q(j)$: there is $\triangle(j) = \frac{q}{p}\triangle(j-1)$, then $\triangle(j) = (\frac{q}{p})^j$, and $\triangle(0) = (\frac{q}{p})^j q(1)$. Thus, $\triangle(j) = r^j q(1)$. Then

$$\sum_0^{n-1} \triangle(j) = q(1)\sum_0^{n-1} r^j = \begin{cases} nq(1), r=1 \Leftrightarrow\ p=\frac{1}{2} \\ q(1)\frac{1-r^n}{1-r},\ r\neq 1 \end{cases}.$$

Hence,

$$q(1) = \begin{cases} \frac{1}{n}, r=1 \Leftrightarrow p=\frac{1}{2} \\ \frac{1-r}{1-r^n}, r\neq 1 \end{cases}.$$

$$q(0) = \sum_0^{j-1} \triangle(i) = q(1)\sum_0^{j-1} r^i$$

$$(i) r = 1; q(1) = \frac{1}{n}; \sum_0^{j-1} r^i = j$$

$$(ii) r \neq 1; q(1) = \frac{1-r}{1-r^n}; \sum_0^{j-1} r^i = \frac{1-r^0}{1-r}$$

$$q(j) = q(1)\sum_0^{j-1} r^i = \frac{1-r}{1-r^n}\frac{1-r^0}{1-r} = \frac{1-r^j}{1-r^n}$$

$$p(0) = \frac{r^j - r^n}{1-r^n}$$

Example 1.3. *For a simple example, let us say $p = 0.4$, then $r = \frac{0.6}{0.4} = 1.5$, for $j = 5$, $n = 8$. Then we have $q(5) = \frac{1-r^5}{1-r^8} = \frac{1.5^5-1}{1.5^8-1} = \frac{1688}{6305} \approx 0.2677$, and $p(5) = 1 - q(5) = \frac{4617}{6305} \approx 0.7323$.*

Consider the following game. For 7 and 11, the player wins. For 2, 3, and 12, the player loses. For the rest (4,5,6,8,9,10), then he/she continues to roll until the 1st time an 8 or 7 appears. If he/she rolls an 8, the player wins. (SOMETHING IS WRONG IN THE NOTES) The goal is to find $Pr(player\ wins)$.

We can construct the following table:

Outcome	Pr(Outcome)	Pr(Win|Outcome)	P(Outcome)
4,10	each $\frac{3}{36}$	$\frac{1}{3}$	$2 \times \frac{1}{36} = \frac{1}{18}$
5,9	each $\frac{4}{36}$	$\frac{2}{5}$	$2 \times \frac{2}{45} = \frac{4}{45}$
6,8	each $\frac{5}{36}$	$\frac{5}{11}$	$2 \times \frac{25}{36(11)} = \frac{25}{187}$
7	$\frac{1}{6}$	1	$\frac{1}{6}$
11	$\frac{1}{18}$	1	$\frac{1}{18}$

Thus we have $Pr(win) = \sum_K Pr(win|Outcome\ k)Pr(k) = \frac{244}{395} \approx 0.4929$.

2 Elements of Probability

Go back to Table of Contents. Please click TOC

We start this section by definitions and explanation of different rules. With intuitive understanding from last section, we establish the concepts with graphical representation.

We have the following rules:

Definition 2.1. We list a few operations of forming unions, intersections, and complements of events.
Commutative Law: $E \cup F = F \sup E$ or $EF = FE$
Associative Law: $(E \cup F) \cup G = E \cup (F \cup G)$ or $(EF)G = E(FG)$
Distributive Law: $(E \cup F)G = EG \cup FG$ or $EF \cup G = (E \cup G)(F \cup G)$.

We can then discuss the following relationship between three basic operations known as DeMorgan's Laws.

$$(E \cup F)^C = E^C F^C$$

$$(EF)^C = E^C \cup F^C$$

Suppose every event E of an experiment having a sample space S there is a number, denoted by $P(E)$, that is in accord with the following three axioms:

(i) $0 \leq P(E) \leq 1$

(ii) $P(S) = 1$

(iii) For any sequence of mutually exclusive events $E_1, E_2, ...$ (that is, events for which $E_i E_j = \emptyset$ when $i \neq j$,

$$P\left(\cup_{i=1}^n E_i \right) = \sum_{i=1}^n P(E_i),\ n = 1, 2, ..., \infty$$

We call $P(E)$ the probability event of E.

Proposition 2.2.

$$P(E^C) = 1 - P(E)$$

Proposition 2.3.

$$P(E \cup F) = P(E) + P(F) - P(EF)$$

2.1 Conditional Probability

Go back to Table of Contents. Please click TOC

Let E and F be events. Then the conditional probability of E given F is defined as ,

$$P(E|F) = \frac{P(EF)}{P(F)}, \ for \ P(F) > 0$$

In special case, we have

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{N(EF)}{N(F)}$$

Example 2.4. *For an example, a goal can be to find the probability that a poker hand contains at least on 3, given that contains at least one 7. Then we simply let F = at least one 7, and E = at least one 3. Both events happen at the same time can be EF = at least one 3 and at least one 7. Then we calculate the following*

$$P(E|F) = \sum_{k=1}^{4} \sum_{l=1}^{5-k} \frac{\binom{4}{k}\binom{4}{l}\binom{44}{5-k-l}}{\binom{52}{5}}$$

However it can be easier to use the following

$$P(EF) = 1 - P(E^C \cap F^C)$$

$$= 1 - [P(E^C) + P(F^C) - P(E^C F^C]$$

$$= 1 - P(E^C) - P(F^C) + P(E^C F^C)$$

$$P(F^C) = P(E^C) = \frac{\binom{48}{5}}{\binom{52}{5}} \approx 0.6588$$

$$P(F^C E^C) = \frac{\binom{44}{5}}{\binom{52}{5}} \approx 0.4179$$

$$P(EF) \approx 0.1002$$

2.2 Law of Total Probability

Go back to Table of Contents. Please click TOC

Let $F_1, ..., F_n$ "partition" the sample space S meaning that $\bigcup_1^n F_j = S$ and the $\{F_j\}$ are non-overlapping, that is, $F_j F_k = \emptyset$ for $j \neq k$. Then

$$P(E) = \sum_{1}^{n} P(EF_i) = \sum_{1}^{n} P(E|F_i)P(F_i)$$

Example 2.5. *For example, a goal could be to find the probability of rolling an 8 with a fair pair of dice. The solution is simple. Let F_i = $\{1st \ die \ equals \ i\}$, $\forall i = 1, ..., 6$. Then $E = \{sum = 8\}$. Then we have $P(F_i) = \frac{1}{6}$, $i = 1, ..., 6$ and $P(E|F_i) = \frac{1}{6}$, $i = 2, 3, 4, 5, 6$ while $P(E|F_1) = 0$. Thus*

$$P(E) = \frac{5}{36}$$

$$P(EF_i) = \begin{cases} \frac{1}{36}, & i=2,...,6 \\ 0, & i=1 \end{cases}$$

2.3 Bayes Law

Go back to Table of Contents. Please click TOC

Let $F_1, ..., F_n$ be a partition of S, then

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(F_j)P(E|F_j)}{\sum_k P(F_k)P(E|F_k)}$$

Example 2.6. *For another example, there are 3 types of flashlights carried by a given store, types A, B, C. The store has 103 of type A, 91 of type B, and 79 of type C. Type A flashlights hav a 70% chance of lasting 10 hours of use; Type B a 60% chance, and Type C a 52% chance (again of lasting 10 hours of use).*

(i) If a random flashlight is chosen, find the probability that it lasts 10 hours.

(ii) A random flashlight is chosen and fails prior to 10 hours of use. Find the probability that it is type B.

The solution is as the following

(i) Let E = the set of events that last 10 hours, A, B, C play the roles of F_1, F_2, and F_3,

$$P(A) = \frac{103}{273}, P(B) = \frac{91}{273}, P(C) = \frac{79}{273}$$

That is,

$$P(E|A) = 0.7, \ P(E|B) = 0.6, P(E|C) = 0.52$$

Then

$$P(E) = P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C)$$

$$= \frac{1}{273}[103(0.7) + 91(0.6) + 79(0.52)] \approx 0.6146$$

(ii) $P(B|E^C) = \frac{P(E^C|B)P(B)}{P(E^C|A)P(A)+P(E^C|B)P(B)+P(E^C|C)P(C)} \approx 0.3459$

Example 2.7. *Let us see another example: find the probability that a poker hand will contain at least two aces, given that it contains at least one ace. Let, A_1 to be {at least one aace}. Let A_2 to be {at least two aces}. Then $A_1A_2 = A_2$, thus*

$$P(A_2|A_1) = \frac{P(A_1A_2)}{P(A_1)} = \frac{P(A_2)}{P(A_1)}.$$

Now we have

$$P(A_1) = 1 - Pr(no\ aces) = 1 - \frac{\binom{48}{5}}{\binom{52}{5}} = \frac{866,656}{\binom{52}{5}},$$

and

$$P(A_2) = P(A_1) - Pr(exactly\ one\ ace) = P(A_1) - \frac{\binom{4}{1}\binom{48}{4}}{\binom{52}{5}} \approx 0.1222$$

2.4 Independence of Events

Go back to Table of Contents. Please click TOC

E and F are defined to be independent if $P(EF) = P(E)P(F)$, when $P(F) > 0$ the definition is equivalent to, $P(E|F) = P(E)$, and when $P(E) > 0$ to $P(F|E) = P(F)$. We prefer the first equation because it does not require that either $P(E)$ or $P(F)$ to be positive.

Example 2.8. *An example can be the following: a set of k coupons, each of which is independently a type j coupon with probability* $P(j)$, $\sum_1^n P(j) = 1$ *is collected. Find the probability that the set contains at least one type j coupon given that it contains at least one type i coupon* $(i \neq j)$. *The solution is straightforward. Define* $A_i = \{$ *at least one type i coupon in sample*$\}$; *similarly define* A_j: $P(A_i) = 1 - P(A_i^C) = 1 - q_i^k$. *Then* $P(A_iA_j) = 1 - P(A_i^C \cap A_j^C) = 1 - [P(A_i^C) + P(A_j^C) - P(A_i^C A_j^C)]$, *which becomes* $P(A_iA_j) = 1 - q^k(i) - q^k(j) + (1 - p(i) - p(j))^k$.

Example 2.9. *A quick numerical example is the following: Two cards are random dealt without replacement. Find,*

(i) Pr(2 spades—1st card is a spade)

(ii) Pr(2 spades—at least one spade)

The solution is straightforward: if the first card is a spade, then there are 12 spades left in the 51 remaining cards. Thus

$$Pr(2\ spades|first\ card\ is\ a\ spade) = \frac{12}{51} \approx 0.2353$$

.

For this one, it is a simple conditional probability as well

$$Pr(2\ spades|at\ least\ one\ spade) = \frac{Pr(2\ spades)}{Pr(at\ least\ one)} = \frac{\frac{1}{4}\frac{4}{17}}{1 - \frac{3}{4}\frac{38}{51}} \approx 0.133$$

3 Random Variables and Expectation

Go back to Table of Contents. Please click TOC

A random variable is a real (as opposed to vector or complex number) valued function defined on the sample space. There may be many random variables of interest on the same sample space. For example, suppose the experiment consists of rolling a die 3 times, with the sample space S consisting of the 216 ordered triplets (i, j, k) with i, j, k each ranging from 1 to 6. Some of the random variables of interest may be,

$$X(i, j, k) = i + j + k,$$

the sum of the rolls

$$Y(i, j, k) = max(i, j, k)$$

$$Z(i, j, k) = \ outcome\ of\ the\ first\ roll\ = i$$

That is, let $X(1,2,3) = 6$, $Y(1,2,3) = 3$, and $Z(1,2,3) = 1$. If X is a random variable defined on a sample space S, with probability distribution P on S, then for A, a subset of numbers:

$$P(X \in A) = \sum_{\{s:X(s)\in A\}} P(s)$$

In the above example, if $P(i,j,k) = \frac{1}{216} \forall (i,j,k) \in S$, then

$$X(i,j,k) = i+j+k : P(X=6) = \frac{10}{216},$$

with sample points $\{s : X(s) = 6\}$ are the 6 permutations of (1,2,3). The 3 sample points with a 4 and two 1's and the sample point (2,2,2). Next, we have

$$Y(i,j,k) = max(i,j,k) : Pr(Y \leq 3) = (\frac{1}{2})^3 = \frac{1}{8}$$

while each of the outcome must fall in $\{1,2,3\}$. Then we have

$$Z(i,j,k) = i : Pr(Z > 1) = \frac{5}{6}$$

which is the first outcome must belong to $\{2,3,4,5,6\}$.

3.1 Discrete Random Variables

Go back to Table of Contents. Please click TOC

Discrete random variables take their values in a set of real number $\{x_i : i = 1, ...\}$ either finite or countably infinite. Often they are integer valued.

The distribution is characterized by its probability mass function (p.m.f.) with

$$P(x_i) \geq 0 \; and \; \sum_i P(x_i) = 1$$

We can look at the following example for illustration. Consider the dice example with $Y(i,j,k) = max(i,j,k)$ and with all 216 sample points equally likely,

$$P(Y=k) = Pr(Y \leq k) - Pr(Y \leq k-1)$$
$$= (\frac{k}{6})^3 - (\frac{k-1}{6})^3 = \frac{2k^2-3k+1}{216}$$

Following this calculation, we can draw the following table

$pmf \; of \; Y$

k	Pr(Y=k)=P(k)
1	$(\frac{1}{6})^3 = \frac{1}{216}$
2	$(\frac{1}{3})^3 - (\frac{1}{6})^3 = \frac{8-1}{216} = \frac{7}{216}$
3	$(\frac{1}{2})^3 - (\frac{1}{3})^3 = \frac{27-8}{216} = \frac{19}{216}$
4	$(\frac{2}{3})^3 - (\frac{1}{2})^3 = \frac{64-27}{216} = \frac{37}{216}$
5	$(\frac{5}{6})^3 - (\frac{2}{3})^3 = \frac{125-64}{216} = \frac{61}{216}$
6	$1-(\frac{5}{6})^3 = \frac{216-125}{216} = \frac{91}{216}$

Remark 3.1. The random variables can be discrete or continuous. We define the probability mass function $p(a)$ of X by $p(a) = P\{X = a\}$. The probability mass function $p(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values $x_1, x_2, ...$, then $p(x_i) > 0$ if $i = 1, 2, ...$ while $p(x) = 0$, for all other values of x. Since X must take on one of the values x_i, we have $\sum_{i=1}^{\infty} p(x_i) = 1$.

Remark 3.2. The cumulative distribution function F can be expressed in terms of $p(x)$ by

$$F(a) = \sum_{all\ x \leq a} p(x).$$

If X is a discrete random variable whose set of possible values are x_1, x_2, x_3, ..., where $x_1 < x_2 < x_3 < ...$, then its distribution function F is a step function. That is, the value of F is constant in the intervals $[x_{i-1}, x_i)$ and then takes a step (or jump) of size $p(x_i)$ at x_i.

Whereas the set of possible values of a discrete random variable is a sequence, we often must consider random variables whose set of possible values is an interval. Let X be such a random variable. We say that X is a continuous random variable if there exists a nonnegative function $f(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)dx$$

The function $f(x)$ is called the probability density function of the random variable X.

The equation above states that the probability that X will be in B may be obtained by integrating the probability density function over the set B. Since X must assume some value, $f(x)$ must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

If we let $B = [a, b]$, we obtain the following

$$P\{a \leq X \leq b\} = \int_a^b f(x)dx$$

If $a = b$, then

$$P\{X = a\} = \int_a^a f(x)dx = 0,$$

which states that the probability that a continuous random variable will assume any particular value is zero.

A natural question to seek answer afterwards is to identify the relationship between cumulative distribution $F(...)$ and the probability density $f(..)$, which is

$$F(a) = P\{X \in (-\infty, a]\} \int_{\infty}^{a} f(x)dx$$

Remark 3.3. Differentiating both sides, we have $\frac{d}{da}F(a) = f(a)$.

3.2 Properties of Expected Values

Go back to Table of Contents. Please click TOC

One of the most important concepts in probability theory is that of the expectation of a random variable. If X is a discrete random variable taking on the possible values x_1, x_2, ..., then the expectation of expected value of X, denoted by $E[X]$, is defined by

$$E[X] = \sum_i x_i P\{X = x_i\}$$

Consider now a random variable X, which must take on one of the values $x_1, ..., x_n$ with respective probabilities $p_1, ..., p_n$. As $log_2(p_i)$ represents the information conveyed by the message that X is equal to x_i, it follows that the expected amount of information that will be conveyed when the value of X is transmitted is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

The quantity $H(X)$ is known in information theory as the entropy of the random variable X.

We can also define the expectation of a continuous random variable. Suppose that X is a continuous random variable with probability density function f. Since, for dx small

$$f(x)dx \approx P\{x < X < x + dx\}$$

It follows that a weighted average of all possible values of X, with the weight given to x equal to the probability that X is near x, is just the integral over all x of $xf(x)dx$. Hence it is natural to define the expected value of X by

$$E[X] = \int_{\infty}^{\infty} xf(x)dx$$

The properties of expected values are as the following. Consider the expected value of some number as the following

$$E(aX + b) = aE(X) + b$$

which is in the form of linearity. Then for special cases:

$$\begin{cases} b = 0, & if\ E(aX) = aE(X) \\ a = 0, & if\ Eb = b \end{cases}$$

with a and b are constants. We need to note the following. If X takes on positive integer values, with $P(j) = \frac{(\frac{1}{j})^2}{\sum_1^{\infty}(\frac{1}{k})^2}$, then since $\sum_1^{\infty} \frac{1}{k^2} < \infty$ (its value is $\frac{\pi^2}{6}$, $\sum_1^{\infty} P_j = 1$. However,

$$E(X) = \sum_j P(j) = \frac{\sum_1^{\infty} \frac{1}{j}}{\sum_1^{\infty} \frac{1}{k^2}} = \infty$$

As the harmonic series, $\sum_1^\infty \frac{1}{j}$ diverges to ∞. Thus, the expected value, $E(\sum_1^n c_i X - i) = \sum_1^n c_i E(x_i)$ should under the condition of $X_1, ..., X_n$ are random variables with finite means then the result would hold.

The interpretation is straightforward. The mean of X is generally not a typical value. For example the average family size is 2.46 children. The basic intuitive interpretation is that if we have a large sample $X_1, ...X_n$ independently identically distributed with the distribution of X, then $\bar{X}_n = \frac{1}{n}\sum_1^n X_i$ should be close to $\mu = E(X)$. This proximity can be quantified and the strong L.L.N. holds $P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$.

Proposition 3.4. *We introduce the following propositions for expectation of a function of a random variable:*

(a) If X is a discrete random variable with probability mass function $p(x)$, then for any real-valued function g,

$$E[g(X)] = \sum_x g(x)p(x)$$

(b) If X is a continuous random variable with probability density function $f(x)$, then for any real-valued function g,

$$E[g(X)] = \int_\infty^\infty g(x)f(x)dx$$

Corollary 3.5. *An immediate corollary of proposition above is the following. If a and b are constants, then*

$$E[aX + b] = aE[X] + b$$

Proof: We show the proof by two cases: discrete and continuous. We first show the proof in discrete case,

$$\begin{aligned} E[aX + b] &= \textstyle\sum_x (ax + b)p(x) \\ &= a\textstyle\sum_x xp(x) + b\sum_x p(x) \\ &= aE[X] + b \end{aligned}$$

Then we show the proof in the continuous case,

$$\begin{aligned} E[aX + b] &= \textstyle\int_\infty^\infty (ax + b)f(x)dx \\ &= a\textstyle\int_{-\infty}^\infty xf(x)dx + b\int_{-\infty}^\infty f(x)dx \\ &= aE[X] + b \end{aligned}$$

Q.E.D.

Remark 3.6. The expected value of a random variable X, $E[X]$, is also referred to as the mean or the first moment of X. The quantity $E[X^n]$, $n \geq 1$, is called the nth moment of X.

3.3 Expected Value of Sums of Random Variables

Go back to Table of Contents. Please click TOC

The two-dimensional version of proposition in the previous subsection states that if X and Y are random variables and g is a function of two variables, then

$$\begin{aligned} E[g(X,Y)] &= \sum_y \sum_x g(x,y)p(x,y), \text{ in the discrete case} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy, \text{ in the continuous case} \end{aligned}$$

3.4 Variance

Go back to Table of Contents. Please click TOC

The variance of X is defined by

$$\sigma^2 = var(X) = E[(X-\mu)^2] = \sum (x_i - \mu)^2 P(x_i)$$

An alternative formula for $Var(X)$ can be derived as follows: in the discrete case. The standard deviation σ is the positive square root of the variance σ^2.

Example 3.7. *For a simply example, we have following information:*

x	$P(x)$	$(x-\mu)^2$
1	*0.2*	*1.21*
2	*0.5*	*0.01*
3	*0.3*	*0.81*

Then we can calculate mean, variance, and standard deviation

$$\mu = 0.2 + 2(0.5) + 3(0.3) = 2.1$$

$$\sigma^2 = 0.2(1.21) + 0.5(0.01) + 0.3(0.81) = 0.49$$

and

$$\sigma = \sqrt{0.49} = 0.7$$

Note that if the above were measurements in $feet$, then variance would be in units of $feet^2$. The variance and stand deviation are measurements of variability of the distribution, which of course are one of approaches of this goal. Another useful formula for calculating variance:

$$\sigma^2 = E(X^2) - (E(X))^2 = \sum x_i^2 P(x_i) - (\sum x_i P(x_i))^2$$

and this can be applied to the previous example with the same results calculated. The equation is derived from perfect square, which is shown in the following

$$\begin{aligned} Var(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - E[2\mu X] + E[\mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - \mu \end{aligned}$$

Remark 3.8. Note that in this case, $E[X] = \mu$ by definition, and that is how we make the second to last step.

We also want to notice the following equation describing variance.

Definition 3.9. If X is a random variable with corresponding probability density function $f(x)$, then we define the expected value of X to be

$$E(X) = \int_{\infty}^{\infty} xf(x)dx$$

We define the variance of X to be

$$Var(X) = \int_{\infty}^{\infty} [x - E(X)]^2 f(x)dx$$

We also have the following alternative for variance:

$$\begin{array}{rcl} Var(X) & = & \int_{\infty}^{\infty}[x - E(X)]^2 f(x)dx \\ & = & \int_{\infty}^{\infty}[x^2 - exE(X) + E(X)^2]f(x)dx \\ & = & \int_{\infty}^{\infty} x^2 f(x)dx - 2E(X)\int_{\infty}^{\infty} xf(x)dx + EX()^2 \int_{\infty}^{\infty} f(x)dx \\ & = & \int_{\infty}^{\infty} x^2 f(x)dx - 2E(X)E(X) + E(X)^2 \times 1 \\ & = & \int_{\infty}^{\infty} x^2 f(x)dx - E(X)^2 \end{array}$$

Remark 3.10. Note that for X and Y to be random variables we need to discuss discrete and continuous cases with means μ_X and μ_Y, respectively.

If X and Y are discrete variables with joint support S, then the covariance of X and Y is

$$Cov(X, Y) = \sum \sum_{(x,y)\in S} (x - \mu_X)(y - \mu_Y)f(x, y)$$

If X and Y are continuous random variables with supports S_1 and S_2, respectively, then the covariance is

$$Cov(X, Y) = \int_{S_2} \int_{S_1} (x - \mu_X)(y - \mu_Y)f(x, y)dxdy$$

Remark 3.11. Note that $\sigma^2 < \infty$ is and only if $E[X]^2 < \infty$. To be more precise, if $var(x_i) < \infty$, $i = 1, ..., n$, then $var(\sum c_i X_i) = \sum c_i^2 var(X_i)$. If X and Y are independent, then $var(X - Y) = var(X + Y) = var(X) + var(Y)$.

Example 3.12. *For example, a fair die is rolled twice. Let $S = X_1 + X_2$ denote the sum of the rolls. Find the variance of S, assuming that X_1 and X_2 are independent. The solution is the following*

$$E(X_1) = \frac{1}{6}\sum_{1}^{6} j = \frac{1}{6}\frac{6(7)}{2} = 3.5$$

$$E(X_1^2) = \frac{1}{6}\sum_{1}^{6} j^2 = \frac{1}{6}\frac{6(7)(13)}{6} = \frac{91}{6}$$

$$var(X_1) = \frac{91}{6} - (\frac{7}{2})^2 = \frac{35}{12}.$$

Thus, $var(S) = var(X_1 + X_2) = 2var(X_1) = \frac{35}{6}$ and $E(S) = E(X_1 + X_2) = 2E(X_1) = 7$. If $X_n = \sum_1^n X_i$, the sum of n rolls, then $E(S_n) = nE(X) = \frac{7n}{2}$, $var(S_n) = nvar(X) = \frac{35n}{12}$.

Note the identity allows us to compute the mean and variance of S_n without working with the distribution of S_n. The alternative for $n = 2$ would be

$$E(S_2) = \sum_{2}^{12} kP(S_2 = k) = \frac{1}{36}[2(1) + 3(2) + 4(3) +$$

$$5(4) + 6(5) + 7(6) + 8(5) + 9(4) + 10(3) + 11(2) + 12(1)]\frac{252}{36} = 7.$$

A useful identity concerning variances is that for any constants a and b,

$$Var(aX + b) = a^2 Var(X)$$

Proof:

$$\begin{aligned} Var(aX + b) &= E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - a\mu - b)^2] \\ &= E[(aX - a\mu)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2 E[(X - \mu)^2] \\ &= a^2 Var(X) \end{aligned}$$

Q.E.D.

Remark 3.13. Analogous to the mean's being the center of gravity of a distribution of mass, the variance represents, in the terminology of mechanics, the moment of inertia.

Notice that $Var(X + X) = Var(2X) = 2^2 Var(X) = 4Var(X) \neq Var(X) + Var(X)$. This is because the equation ignores the fact that there might be covariance between the two variables.

The covariance of two random variables X and Y, written $Cov(X, Y)$, si defined by

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

where μ_x and μ_y are the means of X and Y, respectively. A useful expression for $Cov(X, Y)$ can be expanding the right side of the definition. This yields

$$\begin{aligned} Cov(X, Y) &= E[XY - \mu_x Y - \mu_y X + \mu_x \mu_y] \\ &= E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y \\ &= E[XY] - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Remark 3.14. Another property of covariance, which immediately follows from the definition, is that, for any constant a,

$$Cov(aX, Y) = aCov(X, Y)$$

Lemma 3.15. *Covariance, like expectation, possesses an additive property.* $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$.

Proof:

$$\begin{aligned} &Cov(X+Z,Y) \\ &\quad = E[(X+Z)Y] - E[X+Z]E[Y],\ from\ the\ covariance\ definition \\ &\quad = E[XY] + E[ZY] - (E[X]+E[Z])E[Y] \\ &\quad = E[XY] - E[X]E[Y] + E[ZY] - E[Z]E[Y] \\ &\quad = E[XY] - E[X]E[Y] - E[Z]E[Y] \\ &\quad = Cov(X,Y) + Cov(Z,Y) \end{aligned}$$

Q.E.D.

Remark 3.16. The above lemma can be easily generalized to show that

$$Cov\Big(\sum_{i=1}^{n} X_i, Y\Big) = \sum_{i=1}^{n} Cov(X_i, Y)$$

which gives rise to the following

$$Cov\Big(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

Remark 3.17. In general, it can be shown that a positive value of $Cov(X,Y)$ is an indication that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases. The strength of the relationship between X adn Y is indicated by the correlation between X and Y, a dimensionless quantity obtained by dividing the covariance by the product of the standard deviations of X adn Y. That is,

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

and this quantity is always between -1 and $+1$.

Example 3.18. *For example, consider the following case for the understanding of finding variances and covariance. Let X and Y be independent with*

$$f_X(x) = \begin{cases} 2x & ,0 < x < 1 \\ 0 & ,\ elsewhere \end{cases}$$

$$f_Y(y) = \begin{cases} 3y^2 & ,0 < x < 1 \\ 0 & ,\ elsewhere \end{cases}$$

Define $U = X + Y$ and $V = 2X - 3Y$. Please find $var(U)$, $var(V)$, and $cov(U,V)$.

The solution is the following. We can calculate $var(U)$ and $var(V)$ by definition first.

$$\begin{aligned} var(U) &= var(X) + var(Y) \\ var(V) &= 4var(X) + 9var(Y) \end{aligned}$$

Then we need $var(X)$ and $var(Y)$ to calculate $var(U)$ and $var(V)$. We find them by taking integral of $xf(x)$ and $yf(y)$, respectively.

$$E(X) = \frac{2}{3},\ E(X^2) = \frac{1}{2} \Rightarrow var(X) = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

$$E(Y) = \frac{3}{4},\ E(Y^2) = \frac{3}{5} \Rightarrow var(Y) = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}$$

Then we plug in the formula for U and V to find the variances for them.

$$var(U) = var(X) + var(Y) = \frac{1}{18} + \frac{3}{80} = \frac{3}{80}$$

$$var(V) = 4var(X) - 9var(Y) = 3(\frac{1}{18}) + 9(\frac{3}{80}) = \frac{403}{720}$$

Finally, we can calculate $cov(U, V)$, so we have

$$\begin{aligned} cov(U,V) &= cov(X+Y, 2X-3Y) \\ &= cov(X,2X) + cov(X,-3Y) + cov(Y,2X) + cov(Y,-3Y) \\ &= 2var(X) - 3cov(X,Y) + 2cov(X,Y) - 3var(Y) \\ &= -\tfrac{1}{720} \end{aligned}$$

3.5 Joint Probability Mass Function

Go back to Table of Contents. Please click TOC

This subsection we discuss joint probability mass functions [1]. If X and Y are discrete random variables, then the joint p.m.f. of X and Y is defined as

$$P(x_i, y_i) = Pr(X = x_i, Y = y_i)$$

where $\{x_i\}$, $\{y_i\}$ are the mass points of X and Y. It follows from the law of total probability that

$$P_X(x_i) = P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j P(x_i, y_i)$$

and

$$P_Y(y_j) = P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i P(x_i, y_j)$$

Example 3.19. *For example, in a 5 card poker hand from a standard deck define $X = \{\#3's\}$ and $Y = \{\#5's\}$. Then*

$$P(i,j) = \begin{cases} \frac{\binom{4}{i}\binom{4}{j}\binom{44}{5-i-j}}{\binom{52}{5}},\ 0 \le i \le 4,\ 0 \le j \le 4,\ i+j \le 5 \\ 0,\ elsewhere \end{cases}$$

is the joint probability mass function.

Remark 3.20. To specify the relationship between two random variables, we define the joint cumulative probability distribution function of X and Y by

$$F(x, y) = P\{X \le x, Y \le y\}$$

A knowledge of the joint probability distribution function enables one, at least in theory, to compute the probability of any statement concerning the values of X and Y.

We say that X and Y are jointly continuous if there exists a function $f(x,y)$ defined for all real x and y, having the property that for every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane)

$$P\{X,Y\} \in C\} = \iint_{(x,y)\in C} f(x,y)dxdy$$

The function $f(x,y)$ is called the joint probability density function of X and Y. If A and B are any sets of real numbers, then by defining $C = \{(x,y) : x \in A, y \in B\}$, we see from the equation above that

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)dxdy$$

Since

$$\begin{aligned} F(a,b) &= P\{X \in (-\infty,a], Y \in (-\infty,b]\} \\ &= \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy \end{aligned}$$

follows, upon differential, that

$$f(a,b) = \frac{\partial^2}{\partial a \partial b}F(a,b)$$

3.6 Several Random Variables

Go back to Table of Contents. Please click TOC

The joint p.m.f. of discrete random variables $X_1, ...X_m$ is given by

$$P_{X_1,...,X_m}(x_1,...,x_m) = Pr(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)$$

where x_i ranges through the mass points of X_i, $i = 1, ...m$. We have

$$\sum_{X_1,...,X_m} P_{X_1,...,X_m}(x_1,...,x_m) = 1$$

and

$$\sum_{x_1,...,x_k} P_{X_1,...,X_k,X_{k+1},...,X_m}(x_1,...,x_k,x_{k+1},...,x_m)$$

$$= P_{X_{k+1},...,X_m}(x_{k+1},...,x_m)$$

The formula above hold for arbitrary arbitrary subsets of indices. If $A = \{i_1,...,i_k\}$ is a proper non-empty subset of $\{1,...,n\}$, then by letting $A^C = \{j_1,...,j_{m-k}\}$, we have

$$\sum_{i_1,...,i_k} P(X_{i_1},...,X_{i_k},X_{j_1},...,X_{j_{m-k}})(x_{i_1},...,x_{i_k},x_{j_1},...,x_{j_{m-k}})$$

$$= P_{X_{j_1},...X_{j_{m-k}}}(x_{x_{j_1},...,x_{j_{m-k}}})$$

Let us look at more definitions, $X_1,...,X_m$ are independent if for all $A_1,...A_m$, then we have

$$Pr(X_1 \in A_1,...,X_m \in A_m) = \prod_{i=1}^{m} Pr(X_i \in A_i)$$

For discrete random variables this is equivalent to,

$$P_{X_1,...,X_m}(x_1,...,x_m) = \prod_{j=1}^{m} P_{X_j}(X_j)$$

Under independence, we have

$$E(X_1,...,X_m) = \prod_{j=1}^{m} E(X_j)$$

We have the following example for illustration. A fair die is rolled n times. Let $N = \{\#6's\}$. Find the variance of N. The solution is the following. Define

$$I_j = \begin{cases} 1, \ if \ 6 \ on \ jth \ roll \\ 0, \ if \ otherwise \end{cases}, j = 1,...,n$$

Then $E(I_j^2) = E(I_j) = \frac{1}{6}$, $var(I_j) = E(I_j^2) - (E(I_j))^2 = \frac{1}{6} - \frac{1}{36} = \frac{5}{36}$, and

$$N = \sum_{1}^{n} I_j$$

the variance of a sum of independent random variables is the sum of the variances. In terms of covariance, we have a more general term. If $X_1,...,X_n$ are i.i.d. random variables and $N_A = \{\#X_i \in A\}$. Then $EN = nP(X \in A)$ and $var(N) = nP(X \in A)$.

The covariance between discrete random variables (X, Y) is defined by

$$cov(X,Y) = E[(X - EX)(Y - EY)] = \sum_{ij} (X_i - EX)(y_i - EY) P_{X,Y}(x_i, y_j)$$

An equivalent form is,

$$cov(X,Y) = E(XY) - E(X)E(Y)$$

(i) $Cov(X,Y) = Cov(Y,X)$

(ii) $Cov(X,X) = Var(X)$

(iii) $Cov(aX + b, cY + d) = acCov(X,Y)$, a, b, c, d constant

(iv) If X is independent of Y, then $Cov(X,Y) = 0$

(v) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$. Under independence, $Var(X+Y) = Var(X) + Var(Y)$

(vi) $Var(\sum_1^n X_i) = \sum_1^n Var(X_i) + 2\sum_{1 \le i \le j \le n} Cov(X < i, X_j)$ thus if $X_1,...,X_n$ are independent then

$$Var(\sum_{1}^{n} X_i) = \sum_{1}^{n} (Var(X_i))$$

3.7 Moment Generating Functions

Go back to Table of Contents. Please click TOC

The moment generating function (m.g.f.) of a discrete random variable X is defined by,

$$\phi(t) = E(e^{tX}) = \sum_x e^{tx} P(x)$$

for all t for which the sun converges.

The moment generating function $\phi(t)$ of the random variable X is defined for all values t by

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tX} p(x), & if\ X\ is\ discrete \\ \int_{-\infty}^{\infty} e^{tX} f(x)dx, & if\ X\ is\ continous \end{cases}$$

We call $\phi(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $\phi(t)$. In general, the nth derivative of $\phi(t)$ evaluated at $t = 0$ equals $E[X^n]$; that is,

$$\phi^n(0) = E[X^n],\ n \geq 1$$

An important property of moment generating functions is that the moment generating function of the sum of independent random variables is just the product of the individual moment generating functions. To se this, suppose that X and Y are independent and have moment generating functions $\phi_X(t)$ and $\phi_Y(t)$, respectively. Then $\phi_{X+Y}(t)$, the moment generating function of $X + Y$, is given by $\phi_{X+Y}(t) = E[e^{t(X+Y)}]$

$$\begin{aligned} \phi_{X+Y}(t) &= E[e^{t(X+Y)}] \\ &= E[e^{tX} e^{tY}] \\ &= E[e^{tX}]E[e^{tY}] \\ &= \phi_X(t)\phi_Y(t) \end{aligned}$$

where the next to the last equality follows the above theorem since X and Y, and thus e^{gX} and e^{tY} are independent.

The moment generating function has the following converges.

(i) If $E|X|^r < \infty$, then $E(X)^r = \frac{d^r}{dt^r}\phi|_{t=0} = \phi^{(r)}(0)$, the r^{th} derivative of the m.g.f. at $t = 0$.

(ii) If $X_1, ..., X_m$ are independent, then

$$\phi_{\sum_1^m X_i}(t) = \prod_1^m \phi_{X_i}(t)$$

Thus, if $X_1, ..., X_m$ are i.i.d. with common m.g.f. ϕ, then

$$\phi_{\sum_1^m X_i}(t) = \phi_X^m(t).$$

(iii) Two different distributions cannot have the same m.g.f. (known as uniqueness of m.g.f.).

We can look at the following example for a better understanding. N the Poisson example, let $X_1, ..., X_m$ be independent with $X_i \sim Poisson(\lambda_i)$. Then

$$\sum_1^m X_i \ \sim \ Poisson(\sum_1^m \lambda_i).$$

We can prove this by showing $\phi_{\sum_1^m X_i}(t) = \prod_1^m \phi_{X_i}(t) = \prod_1^m e^{\lambda_i(e^t-1)} = e^{(\sum_1^m \lambda_i)(e^t-1)}$, which is the m.g.f. of $Poisson(\sum_1^m \lambda_i$. By uniqueness of the m.g.f., $\sum_1^m X_i \sim Poisson(\sum_1^m \lambda_i)$.

3.8 Chebyshev's Inequality and the Weak Law of Large Numbers

Go back to Table of Contents. Please click TOC

We start this section by discussing Markov's inequality.

Proposition 3.21. *Markov's Inequality. If X is a random variable that takes only nonnegative values, then for any value $a > 0$*

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Proof: Assume X is continuous with density f, then we have the following

$$\begin{aligned} E[X] &= \int_0^\infty x f(x)dx \\ &= \int_0^a x f(x)dx + \int_a^\infty x f(x)dx \\ &\geq \int_a^\infty x f(x)dx \geq \int_a^\infty a f(x)dx \\ &= a\int_a^\infty f(x)dx = a\int_a^\infty f(x)dx \\ &= aP\{X \geq a\} \end{aligned}$$

Q.E.D.

Remark 3.22. The inequality is non-trivial for $a > \mu$, in which case $\frac{\mu}{a} < 1$.

Proposition 3.23. *Chebyshev's Inequality. If X is a random variable with mean μ and variance σ^2, then for any value $k > 0$*

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

Proof: Since $(X-\mu)^2$ is a nonnegative random variable, we can apply Markov's inequality (with $a = k^2$) to obtain

$$P\{(X-\mu)^2 \geq k^2\} \leq \frac{E[(X-\mu)^2]}{k^2}$$

But since $(X - \mu) \geq k^2$ if and only if $\{X - \mu| \geq k$, then

$$P\{|X - \mu| \geq k\} \leq \frac{E[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

which completes the proof.

Q.E.D.

Remark 3.24. Notice that Chebyshev's Inequality takes three forms.

(1) $Pr(|Y-\mu_Y| \geq a) = Pr(|Y-\mu_Y|^2 \geq a^2) \leq \frac{E(Y-\mu_Y)^2}{a^2} = \frac{\sigma^2}{a^2}$.

(2) $Pr(\mu_Y - a < Y < \mu_Y - a) = Pr(|Y-\mu_Y| < a) \geq 1 - \frac{\sigma_Y^2}{a^2}$.

(3) Consider $a = c\sigma_Y$ with $c > 0$. Then $Pr(|Y - E(Y)| \geq c\sigma_Y) \leq \frac{1}{c^2}$ and its complement $Pr(\mu_Y - c\sigma_Y < Y < \mu_Y + c\sigma_Y) \geq 1 - \frac{1}{c^2}$.

Theorem 3.25. *Weak Law of Large Numbers. Let X_1, X_2, ..., be a sequence of independent and identically distributed random variables, each having mean $E[X_i] = \mu$. Then, for any $\epsilon > 0$,*

$$P\left\{\left|\frac{X_1 + ... + X_n}{n} - \mu\right| > \epsilon\right\} \to 0 \text{ as } n \to \infty$$

Remark 3.26. We can look at the following example for the application of Weak Law of Large Numbers.

Let $X_1, ..., X_n$ be i.i.d. random variables with mean μ and variance σ^2. Then $X_n = \sum_1^n X_i$ has mean μ and variance $n\sigma^2$.

For the sample mean,

$$E(\bar{X}_n) = E[\frac{1}{n}\sum -1^n X_i] = \frac{E[S_n]}{n} = \frac{n\mu}{n} = \mu$$

and

$$var(\bar{X}_n) = var(\frac{S_n}{n}) = \frac{1}{n^2} var(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Then we would also have

$$Pr(|\bar{X}_n - \mu| \geq a) \leq \frac{var(\bar{X}_n)}{a^2} = \frac{\sigma^2}{na^2} \to 0$$

as $n \to \infty$ for all $a > 0$, which is the "weak law of large numbers". It follows that

$$Pr(\mu - a < \bar{X}_n < \mu + a) \geq 1 - \frac{\sigma^2}{na^2} \to 1$$

as $n \to \infty$. Thus the sample mean gets more and more concentrated around the true mean, μ, as $n \to \infty$.

4 Special Random Variables

Go back to Table of Contents. Please click TOC

4.1 Binomial Distributions

Go back to Table of Contents. Please click TOC

Binomial independent trials with common probability p of success are known as Bernoulli trials. Suppose we have n Bernoulli trials, and we let X denote the random number of successes obtained. Then

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k}, \ k = 0, ..., n$$

where $q = 1 - p$. The argument is that there are $\binom{n}{k}$ subsets of size k from the n trials, where the k successes can occur. Each of these gives an n-tuple $(x_1, ..., x_n)$, with k 1's (for success) and n-k 0's (for failures). By independence a designated n-tuple with k 1's and n-k 0's has probability $p^k q^{n-k}$. Thus

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k}.$$

4.2 Poisson Distribution

Go back to Table of Contents. Please click TOC

A random variable X, taking on one of the values 0, 1, 2, ..., is said to be a Poisson random variable with parameter λ, $\lambda > 0$, if its probability mass function is given by

$$P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}, \ i = 0, 1, ...$$

Remark 4.1. The symbol e stands for a constant approximately equal to 2.7183. It is a famous constant in mathematics, named after the Swiss mathematician L. Euler, and it is also the base of the so-called natural logarithm.

Thus we have probability mass function to be

$$\sum_i^\infty p(i) = e^{-\lambda} \sum_0^\infty \lambda^i i! = e^{-\lambda} e^{\lambda} = 1$$

Remark 4.2. The Poisson probability distribution was introduced by S.D. Poisson in a book he wrote dealing with the application of probability theory to lawsuits, criminal trials, and the like. The book, published in 1837, was entitled *Recherches sur la probabilite des jugements en matiere criminelle et en matiere civile*.

As a prelude to determining the mean and variance of a Poisson rnadom variable, we can first determine its moment generating function

$$\begin{aligned} \phi(t) &= E[e^{tX}] \\ &= \textstyle\sum_{i=0}^\infty e^{ti} e^{-\lambda} \lambda^i / i! \\ &= e^{-\lambda} \textstyle\sum_{i=0}^\infty (\lambda e^t)^i / i! \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= exp\{\lambda(e^t - 1)\} \end{aligned}$$

Differentiation yields

$$\begin{aligned} \phi'(t) &= \lambda e^t exp\{\lambda(e^t - 1)\} \\ \phi''(t) &= (\lambda e^t)^2 exp\{\lambda(e^t - 1)\} + \lambda e^t exp\{\lambda(e^t - 1)\} \end{aligned}$$

Evaluating at $t = 0$ gives that

$$\begin{aligned} E[X] &= \phi'(0) = \lambda \\ Var(X) &= \phi''(0) - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

Thus both the mean and the variance of a Poisson random variable are equal to the paramter λ.

We can consider the following application. Suppose that X is a binomial random variable with parameters (n, p) and let $\lambda = np$. Then

$$\begin{aligned} P\{X = i\} &= \frac{n!}{(n-1)!i!} p^i (1-p)^{n-i} \\ &= \frac{n!}{(n-1)!i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)...(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \end{aligned}$$

Now, for n large and p small,

$$(1 - \lambda/n)^n \approx e^{-\lambda} \frac{n(n-1)...(n-i+1)}{n^i} \approx 1, \ (1 - \frac{\lambda}{n})^i \approx 1$$

Hence, for n large and p small,

$$P\{X = i\} \approx e^{-\lambda} \frac{\lambda^i}{i!}$$

Proposition 4.3. *Closure Property. If $X_1, ..., X_n$ are independent with X_i Poisson(λ_i), then $\sum_1^n X_i$ Poisson$(\sum_1^n \lambda_i)$.*

Proposition 4.4. *Unimodality. For the binomial p.m.f.:*

$$\begin{aligned} \frac{P(Bin(n,p)=k+1)}{P(Bin(n,p)=k)} &= \frac{p(k+1)}{p(k)} \\ &= \frac{n!}{(k+1)!(n-k-1)!} \frac{k!(n-k)!}{n!} \frac{p}{q} \\ &= \frac{n-k}{k+1} \frac{p}{q} \end{aligned}$$

It follows that $p(k+1) > p(k)$ fi and only if $k < np - p = (n+1)p - 1$; similarly $p(k+1) < k(k)$ if and only if $k > (n+1)p - 1$. if p is an integer multiple of $\frac{1}{n+1}$, then $p(k) = p(k+1)$ for $k = (n+1)p - 1$.

4.3 Geometric Distribution

Go back to Table of Contents. Please click TOC

Suppose we have a sequence of Bernoulli trials and X represents the trial at which the first success occurs. Then

$$p(k) = p(X = k) = q^{k-1} p, \ x = 1, 2, ...$$

where $q = -1$, p is the success probability and $0 < p < 1$. We call this distribution geometric with probability p.

$$\phi(t) = Ee^{tX} = \frac{pe^t}{1 - qe^t}, \ t < log(\frac{1}{q})$$

$$E(X) = \frac{1}{p}, \ Var(X) = \frac{q}{p}$$

We also have, $\bar{F}(k) = Pr(X > k) = Pr(1st \ k \ trials \ in \ failure) = q^k, k = 0, 1, ...$

4.4 Negative Binomial

Go back to Table of Contents. Please click TOC

The nagative biinomial distribution with parameter (r, p) is the number of trials needed to achieve r successes in Bernoulli trials with probability p of success. Since the waiting time between the $(r-1)st$ and rth success is geometric (p), and these interarrival times between successes are independent, we can represent,

$$Y \sim NB(r,p) = \sum_{1}^{r} X_i$$

with $X_1, ..., X_r$ i.i.d. geometric (p). We have

$$\phi_Y(t) = \phi_X^t(t) = (\frac{pe^t}{1-qe^t})^r, \ t < log(\frac{1}{q})$$

$$E(Y) = rE(X) = \frac{r}{p}, \ Var(Y) = rVar(X) = \frac{rq}{p^2}$$

4.5 Uniform Distribution

Go back to Table of Contents. Please click TOC

A random variable X is said to be uniformly distributed over the interval $[\alpha, \beta]$ if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & , \quad if \ \alpha \leq x \leq \beta \\ 0 & , \quad otherwise \end{cases}$$

4.6 Multinomial Distribution

Go back to Table of Contents. Please click TOC

Suppose that $X_1, ..., X_n$ are i.i.d., each taking on the values $1, 2, ..., m$ with respective probabilities $p_1, ..., p_m$, where $p_j \geq 0$, $\sum_1^m p_j = 1$, $j = 1, ..., m$. Define $N_j = \{\#X_i \ which \ equal \ j\}$, $j = 1, ..., m$. Then the joint distribution of $\{N_1, N_2, ..., N_m\}$ is called multinomial. The parameters are $(n, p_1, ..., p_m)$. The binomial is a special case of the multinomial with $m = 2$, $p_1 = p$ and $p_2 = 1 = 1 - p$. The joint p.m.f. of a multinomial distribution with parameters, $(m, p_1, ..., p_m)$ is given by

$$p(k_1, ..., k_m) = \frac{n!}{\prod_1^m k_j!} \prod_{j=1}^{m} p_j^{kj}, \ \forall k_j \geq 0, \ \sum_{1}^{m} k_j = n$$

4.7 Normal Distribution

Go back to Table of Contents. Please click TOC

A random variable is said to be normally distributed with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is

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

The normal density $f(x)$ is a bell-shaped curve that is symmetric about μ and that attains its maximum value of $1/\sigma\sqrt{2\pi} \approx 0.3999/\sigma$ at $x = \mu$.

Remark 4.5. The normal distribution was introduced by the French mathematician Abraham de Moivre in 1733 and was used by him to approximate probabilities associated with binomial random variables when the binomial parameter n is large. This result was later extended by Laplace and others and is now encompassed in a probability theorem known as the central limit theorem, which gives a theoretical base to the often noted empirical observation that, in practice, many random phenomena obey, at least approximately, a normal probability distribution.

Notice that $E[X] = \mu$ and $var(X) = \sigma^2$. We also have $E[Y] = E[a+bX] = a+bE[X] = a+b\mu$ and the variance $Var(Y) = Var(a+bX) = b^2Var(X) = b^2\sigma^2$.

Now let us systematically discuss normal distribution.

Consider a population of elements. Let $X_1, X_2, ..., X_n$ be a sample of values from this population. The sample mean is defined by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Since the value of the sample mean $\bar{X}$ is determined by the values of the random variables in the sample, it follows that $\bar{X}$ is also a random variable. Its expected value and variance are obtained

$$\begin{aligned} E[\bar{X}] &= E\left[\tfrac{X_1+\cdots+X_n}{n}\right] \\ &= \tfrac{1}{n}(E[X_1] + \cdots + E[X_n]) \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} Var(\bar{X}) &= Var\left(\tfrac{X_1+\cdots+X_n}{n}\right) \\ &= \tfrac{1}{n^2}[Var(X_1) + \cdots + Var(X_n)],\ \textit{by independence} \\ &= \tfrac{n\sigma^2}{n^2} \\ &= \tfrac{\sigma^2}{n} \end{aligned}$$

where μ and σ^2 are the population mean and variance, respectively. Hence, the expected value of the sample mean is the population mean μ whereas its variance is $1/n$ times the population variance.

Standard Normal.

A random variable Z is said to be standard normal if its p.d.f. is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},\ -\infty < z < \infty$$

The general form is the following

Let Z be standard normal and define X by

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

with $-\infty < \mu < \infty, 0 < \sigma < \infty$.

Then we have

$$E(x) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu$$

and

$$Var(X) = Var(\mu + \sigma Z) = \sigma^2 Var(Z) = \sigma^2$$

For p.d.f. of X,

$$f_{\mu+\sigma Z}(x) = \frac{d}{dx} Pr(\mu + \sigma Z \leq x) = \frac{d}{dx} Pr(Z \leq \frac{x-\mu}{\sigma})$$

$$= \frac{1}{\sigma} f_z(\frac{X-\mu}{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

A random variable with p.d.f. given by the equation above is said to be $\mathcal{N}(\mu, \sigma^2)$ distributed, and the distribution is described as nromal with meanμ and variance σ^2 as normal with mean μ and variance σ^2. There is one normal distribution for each pair (μ, σ^2) with $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$.

Moment Generating Function.

For Z, standard normal,

$$\begin{aligned} M_z(t) &= E(e^{tZ}) = \tfrac{1}{\sqrt{2\pi}} \textstyle\int_{-\infty}^{\infty} e^{gz} e^{-z^2/2} dz \\ &= e^{t^2/2} \textstyle\int_{-\infty}^{\infty} \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-t)^2}{2}} dz \\ &= e^{t^2/2} \textstyle\int_{-\infty}^{\infty} f_{N(t,1)}(z) dz \\ &= e^{t^2/2}(1) = e^{t^2/2} \end{aligned}$$

For $X \sim \mathcal{N}(\mu, \sigma^2) \sim \mu + \sigma z$, we have

$$\begin{aligned} M_x(t) &= E(e^{tX}) = E(e^{t(\mu+\sigma z)}) = e^{t\mu} E(e^{t\sigma z}) \\ &= e^{t\mu} M_z(\sigma t) = e^{t\mu} e^{\frac{\sigma^2 t^2}{2}} \\ &= e^{t\mu + \frac{\sigma^2 t^2}{2}} \end{aligned}$$

Closure Property.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Consequence.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. This property allows replacement calculations involving $\mathcal{N}(\mu, \sigma^2)$ into analogous calculations for $Z \sim \mathcal{N}(\mu, \sigma^2)$ is known as the "z transformation".

Properties of $\mathcal{N}(0, 1)$.

Define $\Phi(t) = Pr(Z \leq t)$ where $Z \sim \mathcal{N}(0, 1)$. Φ is called the standard normal c.d.f.,

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2} dx$$

Note that Φ cannot be handled analytically (closed from formula for calculating $\Phi(t)$), but has been extensively noted in the table.

Some Properties.

Since $-z \sim z$, then $\Phi(t) = Pr(-z \leq t) = Pr(z \geq -t) = 1 - \Phi(-t)$. Thus, we have

$$\Phi(t) = 1 - \Phi(-t), \text{ and } \Phi(-t) = 1 - \Phi(t)$$

Percentitiles.

Define γ_p by $\Phi(\gamma_p) = p$, known as the pth percentile of $\mathcal{N}(0,1)$. Since

$$\Phi(-\gamma_p) = 1 - \Phi(\gamma_p) = 1 - p$$

wee see that $\gamma_{1-p} = -\gamma_p$

The Central Limit Theorem.

Theorem 4.6. *Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of $X_1 + \cdots + X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.*

It follows from the central limit theorem that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal random variable; thus, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

where Z is a standard normal random variable.

Let $X_1, ..., X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean

$$\bar{X} = \sum_{i=1}^{n} X_i/n$$

Since a constant multiple of a normal random variable is also normal, it follows from the central limit theorem that $\bar{X}$ will be approximately normal when the sample size n is large. Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, then follows that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Sample Variance.

Let $X_1, ..., X_n$ be a random sample from a distribution with mean μ and variance σ^2. Let $\bar{X}$ be the sample mean.

Definition 4.7. The statistic S^2, defined by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

is called the sample variance. $S = \sqrt{S^2}$ is called the sample standard deviation.

To compute $E[S^2]$, we use an indentity: for any numbers $x_1, ..., x_n$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$. It follows that

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Taking expectations of both sides of the preceding yields, upon using the fact that for any random variable W, $E[W^2] = Var(W) + (E[W])^2$,

$$(n-1)E[S^2] \quad = \quad E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\bar{X}^2]$$

Sampling Distributions from a Normal Population.

Let $X_1, ..., X_n$ be a sample from a normal population having mean μ and variance σ^2. That is, they are independent and $X_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, ..., n$. Also let

$$\bar{X} = \sum_{i=1}^{n} X_i/n$$

and

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Since the sum of independent normal random variables is normally distributed, it follows that $\bar{X}$ is normal with mean

$$E[\bar{X}] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \mu$$

and variance

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \sigma^2/n$$

That is, $\bar{X}$, the average of the sample, is normal with a mean equal to the population mean but with a variance reduced by a factor of $1/n$. It follows that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable.

Joint Distribution of $\bar{X}$ and X^2.

We not only obtain the distribution of the sample variance S^2, but we also discover a fundamental fact about normal samples, that $\bar{X}$ and S^2 are independent with $(n-1)S^2/\sigma^2$ having a chi-square distribution with $n-1$ degrees of freedom.

To start, for numbers $x_1, ..., x_n$, let $y_i = x_i - \mu, i = 1, ..., n$. Then as $\bar{y} = \bar{x} - \mu$, it follows from the identity

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2.$$

Now, if $X_1, ..., X - n$ is a sample from a normal population having mean μ variance σ^2, then we obtain from the preceding identity that

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}(X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}$$

or, equivalently,

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} + \left[\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}\right]^2.$$

Theorem 4.8. *If $X_1, ..., X_n$ is a sample from a normal population having mean μ and variance σ^2, then $\bar{X}$ and S^2 are independent random variables, with $\bar{X}$ being normal with mean μ and variance σ^2/n and $(n-1)/S^2/\sigma^2$ being chi-square with $n-1$ degrees of freedom.*

Corollary 4.9. *Let $X_i, ..., X_n$ be a sample fro ma normal population with mean μ. If $\bar{X}$ denotes the sample mean and S the sample standard deviation, then*

$$\sqrt{n}\frac{(\bar{X} - \mu)}{S} \sim t_{n-1}$$

Remark 4.10. X is a hypergeometric random variable; and so the preceding shows that a hypergeometric can be approximated by a binomial random variable when the number chosen is small in relation to the total number of elements.

For mean and standard deviation of a binomial random variable, we see that

$$E[X] = np \text{ and } SD(X) = \sqrt{np(1-p)}$$

Since $\bar{X}$, the proportion of the sample that has the chracteristic, is equal to X/n, we see from the preceding that

$$E[\bar{X}] = E[X]/n = p$$

5 Distribution of Sampling Statistics

Go back to Table of Contents. Please click TOC

Definition 5.1. *Bernoulli Distribution.* A random variable X has Bernoulli distribution with parameter p $(0 \leq p \leq 1)$ if X can take only the values 0 and 1 and the probabilities are

$$Pr(X = 1) = p \text{ and } \Pr(X = 0) = 1 - p.$$

The p.f. of X can be written as follows:

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & , \quad for\ x = 0, 1 \\ 0 & , \quad otherwise \end{cases}$$

If X has the Bernoulli distribution with parameter p, then X^2 and X are the same random variable. It follows that

$$\begin{aligned} E(X) &= 1 \times p + 0 \times (1-p) = p \\ E(X^2) &= E(X) = p \\ &\quad and \\ Var(X) &= E(X^2) - [E(X)]^2 = p(1-p) \end{aligned}$$

Moreover, the m.g.f. of X is

$$\begin{aligned}\phi(t) &= E(e^{tX}) \\ &= pe^t + (1-p) \; for \; -\infty < t < \infty\end{aligned}$$

Definition 5.2. *Binomial Distribution.* A random variable X has the binomial distribution with parameters n and p if X has a discrete distribution for which the p.f. is as follows:

$$f(x|n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & , \quad for \; x = 0,1,2,...,n \\ 0 & , \quad otherwise \end{cases}$$

Theorem 5.3. *If the random variables $X_1, ..., X_n$ from n Bernoulli trials with parameter p, and if $X = X_1 + ... + X_n$, then X has the binomial distribution with parameters n and p. Then we have*

$$\begin{aligned} E(X) &= \sum_{i=1}^n E(X_i) = np \\ Var(X) &= \sum_{i=1}^n Var(X_i) = np(1-p) \\ \phi(t) &= E(e^{tX}) \\ &= \prod_{i=1}^n E(e^{tX_i}) \\ &= (pe^t + 1 - p)^n \end{aligned}$$

Remark 5.4. What about the theorem of Closeness of Binomial and hypergeometric distributions on page 284?

Definition 5.5. *Poisson Distribution.* Let $\lambda > 0$. A random variable X has the Poisson distribution with mean λ if the p.f. of X is as follows:

$$f(x|\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & , \quad for x = 0,1,2,... \\ 0 & , \quad otherwise \end{cases}$$

Theorem 5.6. Mean. *The mean of the distribution with p.f. equal to the above equation is λ.*

Theorem 5.7. Variance. *The variance of the Poisson distribution with mean λ is also λ.*

Theorem 5.8. Moment Generating Function. *The m.g.f. of the Poisson distribution with mean λ is*

$$\phi(t) = e^{\lambda(e^t - 1)}.$$

Definition 5.9. *Moment Generating Function.* If X has the negative binomial distribution with parameters r and p, then the m.g.f. of X is as follows:

$$\phi(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r \; for \; t < \log\left(\frac{1}{1-p}\right).$$

Theorem 5.10. Mean and Variance. *If X has the negative binomial distribution with parameters r and p, the mean and the variance of X must be*

$$E(X) = \frac{r(1-p)}{p} \; and \; Var(X) = \frac{r(1-p)}{p^2}.$$

Remark 5.11. The mean and variance of the geometric distribution with parameter p are the special case of the equation above with $r = 1$.

Definition 5.12. *Normal Distributions.* A random variable X has the normal distribution with mean μ and variance σ^2 ($-\infty < \mu < \infty$ *and* $\sigma > 0$) if X has a continuous distribution with the following p.d.f.

$$f(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma} exp\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big],$$
$$for -\infty < x < \infty.$$

Definition 5.13. *Lognormal Distribution.* If $\log(X)$ has the normal distribution with mean μ and variance σ^2, we say that X has the lognormal distribution with parameters μ and σ^2.

Definition 5.14. *Gamma Function.* For each positive number α, let the value $\Gamma(\alpha)$ be defined by the following integral

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}$$

The function Γ defined above for $\alpha > 0$ is called the gamma function.

Definition 5.15. *Gamma Distributions.* Let α and β be positive numbers. A random variable X has the gamma distribution with parameters α and β if X has a continuous distribution for which the p.d.f. is

$$f(x|\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & , \quad for\ x > 0 \\ 0 & , \quad for\ x \le 0 \end{cases}$$

Theorem 5.16. Moments. *Let X have the gamma distribution with parameters α and β. For $k = 1, 2, ...,$*

$$E(X^k) = \frac{\Gamma(\alpha+k)}{\beta^k\Gamma(\alpha)} = \frac{\alpha(\alpha+1)...(\alpha+k-1)}{\beta^k}.$$

In particular, $E(X) = \frac{\alpha}{\beta}$, and $Var(X) = \frac{\alpha}{\beta^2}$.

Theorem 5.17. Moment Generating Function. *Let X have the gamma distribution with parameters α and β. The m.g.f. of X is*

$$\phi(t) = \Big(\frac{\beta}{\beta-t}\Big)^\alpha \ for\ t < \beta.$$

Definition 5.18. *Beta Function.* For each positive α and β, define

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

The function B is called the beta function.

Theorem 5.19. *For all $\alpha, \beta > 0$,*

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Definition 5.20. *Beta Distributions.* Let $\alpha, \beta > 0$ and let X be a random variable with p.d.f.

$$f(x|\alpha,\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & , \quad for\ 0 < x < 1, \\ 0 & , \quad otherwise. \end{cases}$$

Theorem 5.21. *Suppose that P has the beta distribution with parameters α and β, and the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p. Then the conditional distribution of P given $X = x$ is the beta distribution with parameters $\alpha+x$ and $\beta+n-x$.*

Theorem 5.22. Moments. *Suppose that X has the beta distribution with parameters α and β. Then for each positive integer k,*

$$E(X^k) = \frac{\alpha(\alpha+1)...(\alpha+k-1)}{(\alpha+\beta)(\alpha+\beta+1)...(\alpha+\beta+k-1)}.$$

In particular,

$$E(X) = \frac{\alpha}{\alpha+\beta},$$

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Definition 5.23. *Multinomial Distribution.* A discrete random vector $\mathbf{X} = (X_1, ..., X_k)$whose p.f. is given as

$$f(\mathbf{X}|n, \mathbf{p}) = Pr(\mathbf{X} = x) = Pr(X_1 = x_1, ..., X_k = k)$$

$$= \begin{cases} \binom{n}{x_1,...,x_k} p_1^{x_1}..p_k^{x_k} & , \quad x_1 + ... + x_k = n \\ 0 & , \quad otherwise \end{cases}$$

and has the multinomial distribution with parameters n and $\mathbf{p} = (p_1, ..., -+ p_k)$.

Theorem 5.24. Means, Variance, and Covariance. *Let the random vector $\boldsymbol{X}$ have the multinomial distribution with parameters n and $\boldsymbol{P}$. The means and variances of the coordinates of X are*

$$E(X_i) = np_i \text{ and } Var(X_i) = np_i(1-p_i) \text{ for } i = 1, ..., k.$$

Also, the covariances between the coordinates are

$$Cov(X_i, X_j) = -np_ip_j.$$

6 Parameter Estimation

Go back to Table of Contents. Please click TOC

6.1 Interval Estimates

Go back to Table of Contents. Please click TOC

Suppose that $X_1, ...X_n$ is a sample from a normal population having unknown mean μ and known variance σ^2. It has been shown that $\bar{X} = \sum_{i=1}^{n} X_i/n$ is the maximum likelihood estimator for μ. However, we do not expect that the sample mean $\bar{X}$ will exactly equal μ, but rather that it will be "close". Hence, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within.

Let us say the following. The point estimator $\bar{X}$ is normal with mean μ and variance σ^2/n, thus it follows that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}$$

has a standard normal distribution. Therefore,

$$P\left\{ -1.96 < \sqrt{n}\frac{(\bar{X}-\mu)}{\sigma} < 1.96 \right\} = 0.95$$

or

$$P\left\{ \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \right\} = 0.95$$

That is, "with 95 percent confidence" we assert that the true mean lies within $.196\sigma/\sqrt{n}$ of the observed sample mean. The interval

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \right)$$

is called a 95 percent confidence interval estimate of μ.

6.1.1 C.I.s for a Normal Mean with Unknown Variance

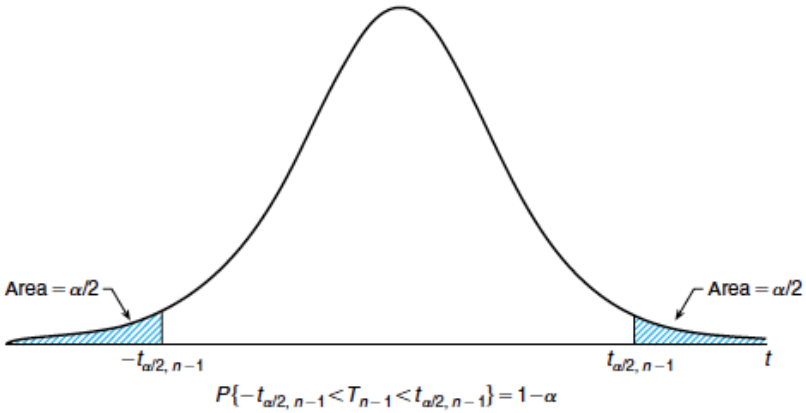
Go back to Table of Contents. Please click TOC

Suppose now that $X_1, ..., X_n$ is a sample from a normal distribution with unknown mean μ and unknown variance σ^2, and that we wish to construct a $100(1-\alpha)$ percent confidence interval for μ. Since σ is unknown, we can no longer base our interval on the fact that $\sqrt{nn}(\bar{X}-\mu)\sigma$ is a standard normal random variable. However, by letting $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ denote the sample variance, then from Corollary it follows that

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S}$$

is a t-random variable with $n-1$ degrees of freedom. Hence, from the symmetry of the t-density function, we have that for any $\alpha \in (0, 1/2)$,

Figure 1: Graphic illustration for t-density function.



Remark 6.1. Recall Corollary 6.5.2. from text [1]. Let $X_i, ..., X_n$ be a sample from a normal population with mean μ. If $\bar{X}$ denotes the sample mean and S the sample standard deviation, then

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1}$$

That is, $\sqrt{n}(\bar{X}-\mu)/S$ has a t-distribution with $n-1$ degrees of freedom.

A one-sided upper vonfidence interval can be obtained by noting that

$$P\left\{\sqrt{n}\frac{(\bar{X}-\mu)}{S} < t_{\alpha,n-1}\right\} = 1-\alpha$$

which is equivalent as

$$P\left\{\mu > \bar{X} - \frac{S}{\sqrt{n}}t_{\alpha,n-1}\right\} = 1-\alpha$$

Hence, if it is observed that $\bar{X} = \bar{x}$, $S = s$¡ then we can assert "with $100(1-\alpha)$ percent confidence" that

$$\mu \in \left(\bar{x} - \frac{s}{\sqrt{n}}t_{\alpha,n-1}, \infty\right)$$

and similarly we say a $100(1-\alpha)$ lower confidence interval would be

$$\mu \in \left(-\infty, \bar{x} + \frac{s}{\sqrt{n}}t_{\alpha,n-1}\right)$$

6.1.2 C.I.s for the Variance of a Normal Distribution

Go back to Table of Contents. Please click TOC

If $X_1, ..., X_n$ is a sample from a normal distribution having unknown parameters μ and σ^2, then we can construct a confidence interval for σ^2 by using the fact that

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hence,

$$P\left\{\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha$$

or, equivalently,

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha$$

Hence, when $S^2 = s^2$, a $100(1-\alpha)$ percent confidence interval for σ^2 is

$$\left\{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right\}.$$

Now we can conclude the following. We say there is $100(1-\alpha)$ percent confidence intervals for

$$X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$$

with

$$\bar{X} = \sum_{i=1}^{n} x_i/n$$

and

$$s = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)},$$

thus we have the following intervals under different assumptions ("k." denotes known, and "u." denotes unknown):

As.	**Para.**	**C.I.**	**Lower Interval**	**Upper Interval**
σ^2 (k.)	μ	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$(-\infty, \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}})$	$(\bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty)$
σ^2 (u.)	μ	$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$	$(-\infty, \bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}})$	$(\bar{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty)$
μ (u.)	σ^2	$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$	$\left(0, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right)$	$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \infty\right)$

6.2 Estimating the Difference in Means of Two Normal Populations

Go back to Table of Contents. Please click TOC

Let $X_1, ..., X_n$ be a sample of size n from a normal population having mean μ_1 and variance σ_1^2 and let $Y_1, ..., Y_m$ be a sample of size m from a different normal population having mean μ_2 and variance σ_2^2 and suppose that the two samples are independent of each other. We are interested in estimating $\mu_1 - \mu_2$.

To obtain a confidence interval estimator, we need the distribution of $\bar{X} - \bar{Y}$. Because

$$\bar{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n)$$
$$\bar{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2/m)$$

follow from the fact that the sum of independent normal random variables is also normal, that

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left((\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

Hence, assuming σ_1^2 and σ_2^2 are known, we have that

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1)$$

and so

$$P\left\{ -z_{\alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} < z_{\alpha/2} \right\} = 1 - \alpha$$

or, equivalently,

$$P\left\{ \bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right\} = 1 - \alpha$$

Hence, if $\bar{X}$ and $\bar{Y}$ are observed to equal $\bar{x}$ and $\bar{y}$, respectively, then a $100(1 - \alpha)$ two-sided confidence interval estimate for $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)$$

One-sided confidence intervals for $\mu_1 - \mu_2$ are obtained in a similar fashion, and a $100(1 - \alpha)$ percent one-sided interval is given by

$$\mu_1 - \mu_2 \in \left(-\infty, \bar{x} - \bar{y} + z_\alpha \sqrt{\sigma_1^2/n + \sigma_2^2/m} \right)$$

6.3 Approximate Confidence Interval for the Mean of a Bernoulli Random Variable

Go back to Table of Contents. Please click TOC

Consider a population of items, each of which independently meets certain standards with some unknown probability p. If n of these items are tested to determine whether they meet the standards, how can we use the resulting data to obtain a confidence interval for p?

If we let X denote the number of the n items that meet the standards, then X is a binomial random variable with parameters n and p. Thus, when n is large, it follows by the normal approximation to the binomial that X is approximately normally distributed with mean np and variance $np(1 - p)$. Hence,

$$\frac{X - np}{\sqrt{np(1 - p)}} \dot{\sim} \mathcal{N}(0, 1)$$

where $\dot{\sim}$ means "is approximately distributed as". Therefore, for any $\alpha \in (0, 1)$,

$$P\{-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1 - p)}} < z_{\alpha/2}\} \approx 1 - \alpha$$

and so if X is observed to equal x, then an approximate $100(1-\alpha)$ percent confidence region for p is

$$\left\{p : -z_{\alpha/2} < \frac{x-np}{\sqrt{np(1-p)}} < z_{\alpha/2}\right\}$$

The foregoing region, however, is not an interval. To obtain a confidence interval for p, let $\hat{p} = X/n$ be the fraction of the items that meet the standards.

We conclude the following. A $100(1-\sigma)$ per cent confidence intervals for $\mu_1 - \mu_2$, for

$$X_1, ..., X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_1, ..., Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2),$$

we have

$$\bar{X} = \sum_{i=1} n\frac{X_i}{n}, S_1^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$
$$\bar{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}, S_2^2 = \sum_{i=1}^{m} \frac{(Y_i - \bar{Y})^2}{m-1}$$

Assumption	**Confidence Interval**
σ_1, σ_2 (k.)	$\bar{X} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$
σ_1, σ_2 (u.) but equal	$\bar{X} - \bar{Y} \pm t_{\alpha/2, n+m-2}\sqrt{(\frac{1}{n} + \frac{1}{m})\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}$
Assumption	**Lower C.I.**
σ_1, σ_2 (k.)	$(-\infty, \bar{X} - \bar{Y} + z_\alpha\sqrt{\sigma_1^2/n + \sigma_2^2/m})$
σ_1, σ_2 (u.) but equal	$(-\infty, \bar{X} - \bar{Y} + t_{\alpha, n+m-2}\sqrt{(\frac{1}{n} + \frac{1}{m})\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}})$

Remark 6.2. Note that a $100(1-\alpha)$ percent confidence interval for p will be of approximate length b when the sample size is

$$n = \frac{(2z_{\alpha/2})^2}{b^2} p(1-p).$$

Now it is easily shown that the function $g(p) = p(1-p)$ attains its maximum value of $\frac{1}{4}$, in the interval $0 \leq p \leq 1$, when $p = \frac{1}{2}$. Thus an upper bound on n is

$$n \leq \frac{(z_{\alpha/2})^2}{b^2}$$

and so by choosing a sample whose size is at least as large as $(z_{\alpha/2})^2/b^2$, one can be assured of obtaining a confidence interval of length no greater than b without need of any additional sampling.

Approximate $100(1-\alpha)$ percent confident intervals for p. X is a $Binomial(n, p)$ Random Variable $\hat{p} = X/n$.

Type of Interval	**Confidence Interval**
Two-sided	$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
One-sided lower	$(-\infty, \hat{p} + z_\alpha\sqrt{\hat{p}(1-\hat{p})/n})$
One-sided upper	$(\hat{p} - z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}, \infty)$

6.4 Evaluating a Point Estimator

Go back to Table of Contents. Please click TOC

Let $\mathbf{X} = (X_1, ..., X_n)$ be a sample from a population whose distribution is specified up to an unknown parameter θ, and $d = d(\mathbf{X})$ be an estimator of θ. How are we to determine its worth as an estimator of θ? One way is to consider the square of the difference between $d(\mathbf{X})$ and θ. However, since $(d(\mathbf{X}) - \theta)^2$ is a random variable, let us agree to consider $r(d, \theta)$, the mean square error of the estimator d, which is defined by

$$r(d, \theta) = \mathbb{E}[(d(\mathbf{X}) - \theta)^2]$$

as an indication of the worth of d as an estimator of θ.

It would be nice if there were a single estimator d that minimized $r(d, \theta)$ for all possible values of θ. However, except in trivial situations, this will never be the case.

Definition 6.3. Let $d = d(\mathbf{X})$ be an estimator of the parameter θ. Then

$$b_\theta(d) = \mathbb{E}[d(\mathbf{X})] - \theta$$

is called the bias of d as an estimator of θ. If $b_\theta(d) = 0$ for all θ, then we say that d is an unbiased estimator of θ. In other words, an estimator is unbiased if its expected value always equals the value of the parameter it is attempting to estimate.

Combining Independent Unbiased Estimators. Let d_1 and d_2 denote independent unbiased estimators of θ, having known variances σ_1^2 and σ_2^2. That is, for $i = 1, 2$

$$E[d_i] = \theta, Var(d_i) = \sigma_i^2$$

Any estimator of the form

$$d = \lambda d_1 + (1 - \lambda) d_2$$

will also be unbiased. To determine the value of λ that results in d having the smallest possible mean square error, note that

$$\begin{aligned} r(d, \theta) &= Var(d) \\ &= \lambda^2 Var(d_1) + (1 - \lambda)^2 Var(d_2) \\ &\quad \textit{by the independence of } d_1 \textit{ and } d_2 \\ &= \lambda^2 \sigma_1^2 + (1 - \lambda)^2 \sigma_2^2 \end{aligned}$$

Differentiation yields that

$$\frac{d}{d\lambda} r(d, \theta) = 2\lambda\sigma_1^2 - 2(1 - \lambda)\sigma_2^2$$

To determine the value of λ that minimizes $r(d, \theta)$,call it $\hat{\lambda}$, set this equal to 0 and solve for λ to obtain

$$2\hat{\lambda}\sigma_1^2 = 2(1 - \hat{\lambda})\sigma_2^2$$

or

$$\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}$$

In words, the optimal weight to give an estimator is inversely proportional to its variance (when all the estimators are unbiased and independent).

7 Hypothesis Testing

Go back to Table of Contents. Please click TOC

The case we considered previously is known as "simple vs. simple", meaning that both H_0 and H_1 contain a single distribution. Now we begin to move to more realistic cases. An important clue is provided by example (from 11/21 Lecture - Hypothesis Testing I). We tested $H_0 : N(0,1)$ versus $H_1 : N(1,1)$. The most powerful level-two test rejects H_0 if $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i > \frac{1}{\sqrt{n}} z_\alpha$. Suppose that we switch the alternative H_1 to $N(\mu, 1)$ where $|mu$ is a fixed number exceeding zero. Then

$$\lambda(\underline{x}) = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_i)^2}}{e^{-\frac{1}{2}x_i^2}} = e^{n\mu\bar{x}-\frac{1}{2}n\mu^2}$$

Since λ is an increasing function of $\bar{x}$, rejecting H_0 for λ large is equivalent as rejecting H_0 for $\bar{X}$ large. Our most powerful level α test again rejects H_0 for $\bar{x} > \frac{z_\alpha}{\sqrt{n}}$. The test is independent of $\mu > 0$; the power does depend on μ. Thus, if we consider the new hypothesis testing problem, we have

$$H_0 : \mu = 0, \ against \ H_1 : \mu > 0$$

The test rejects H_0 if $\bar{x} > \frac{z_\alpha}{\sqrt{n}}$, simultaneously maximizes the power at every $\mu > 0$. We call this test a *uniformly most powerful* level-α test.

Similarly in this normal case: we have

$$H_0 : \mu = \mu_0 \ versus \ H_1 : \mu = \mu_1$$

with $\mu_1 > \mu_0$ (both fixed) and with common variable σ_0^2,

$$\begin{aligned} \lambda(\underline{x}) &= \frac{e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu_1)^2}}{e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}} \\ &= e^{\frac{n}{\sigma_0^2}[(\mu_1-\mu_0)\bar{x}-\frac{1}{2}(\mu_1^2-\mu_0^2)]} \end{aligned}$$

which once again is an increasing function of $\bar{x}$. Thus a U.M.P. level-α test of

$$H_0 : \mu = \mu_0, \ versus \ H_1 : \mu > \mu_0$$

is rejected H_0 if $\bar{x} > \mu_0 + \frac{\sigma_0}{\sqrt{n}} z_\alpha$. For testing,

$$H_0 : \mu = \mu_0 \ versus \ H_1 : \mu < \mu_0, \ with\sigma = \sigma_0$$

$\lambda(\underline{x})$ is a monotonically decreasing function of $\bar{x}$. Thus rejecting for λ large is equivalent as rejecting for $\bar{x}$ small. The U.M.P. level-α test rejects H_0 for

$$\bar{x} < \mu_0 - \frac{\sigma_0}{\sqrt{n}} z_\alpha$$

7.1 U.M.P. One-sided Tests

Go back to Table of Contents. Please click TOC

The ideas of above example can be generalized. Suppose that we have the "monotone likelihood ratio" property. We have a parametric family, say $\{f(x|\theta), \theta \in \Omega\}$ while θ is a real-valued parameter. The M.L.R. property is that for any $\theta_0 < \theta_1$, we have

$$\lambda(\underline{x}) = \prod_j n \frac{f(x_i|\theta_1)}{f(x_i|\theta_0)} = g_{\theta_0,\theta_1}(T(x_1, ..., x_n))$$

where g is an increasing function of T. Then rejecting for λ large is equivalent as rejecting for T large. The U.M.P. level-α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ rejects H_0 for $T > c$, where $P_{\theta_0}(T > c) = \alpha$. In the discrete case, we may need to settle for a U.M.P. approximate level-α test as we will illustrate.

Similarly under M.L.R. (monotone likelihood ratio), the U.M.P. level-α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$, rejects H_0 for $T < c$, where $P_{\theta_0}(T < c) = \alpha$.

If $\lambda(\underline{x})$ is a monotonically decreasing function of $T(x_1, ..., x_n)$ for all $\theta_0 < \theta_1$, then for $\theta = \theta_0$ versus $\theta > \theta_0$, we reject H_0 for $T < c$, and for $\theta = \theta_0$ versus $\theta < \theta_0$, we reject H_0 for $T > c$. We again get U.M.P. level-α tests.

Example 7.1. *Consider $\mathcal{N}(\mu, \sigma_0^2)$ previously considered. We have M.L.R. in $\bar{X}$.*

Example 7.2. *Suppose we have Exponential(θ) for $\theta_0 < \theta_1$, then we have*

$$\lambda(\underline{x}) = \prod_1^n \frac{f(x_1, ..., x_n|\theta_1)}{f(x_1, ..., x_n|\theta_0)} = \left(\frac{\theta_1}{\theta_0}^n e^{-(\theta_1-\theta_0)\sum_1^n x_i}\right)$$

Here λ is a decreasing function of $\sum_1^n x_i$. Thus, the U.M.P. level-α test of $H_0 : \theta = \theta_0$, versus $H_1 : \theta > \theta_0$ rejects H_0 for $s_n = \sum_1^n x_i < c$ where $P_{\theta_0}(S_n < c) = \alpha$. Since, $2\theta_0 s_n \sim \chi^2_{2n}$ under H_0. Out U.M.P. level-α rejects H_0 for $s_n < \frac{1}{2\theta_0}\chi^2_{2n,1-d}$.

Example 7.3. *Geometric (θ), for $\theta_0 < \theta_1$, $s_n = \sum_1^n x_i$, then*

$$\begin{aligned} \lambda(\underline{x}) &= \frac{(1-\theta_1)^{\sum(x_i-1)}\theta_1^n}{(1-\theta_0)^{\sum(x_i-1)}\theta_0^n} \\ &= (\tfrac{\theta_1}{\theta_0})^n (\tfrac{1-\theta_1}{1-\theta_0})^{S_n} (\tfrac{1-\theta_0}{1-\theta_1})^n \end{aligned}$$

For $\theta_0 < \theta_1$, $1 - \theta_0 > 1 - \theta_1$ and thus $\frac{1-\theta_1}{1-\theta_0} < 1$; thus λ is a decreasing function of s_n.

For testing, $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$, the U.M.P. approximate level-α test will reject H_0 for $s_n \leq K$ where k is chosen so that $P_{G_0}(s_n \leq K) \approx \alpha$.

This test will be a U.M.P. level $\alpha^* = P_{\theta_)}(s_n \leq K)$ test, with α^* close to α. Under H_0, $s_n \sim Negative\ Binomial(n, G_0) \approx \mathcal{N}(\frac{n}{\theta_0}, \frac{n(1-G_0)}{\theta_0^2})$. Thus,

$P_{H_0}(S_n \leq K) \approx \Phi(\frac{K+\frac{1}{2}-\frac{n}{\theta}}{\sqrt{\frac{n(1-G_0}{G_0^2}}}) = \Phi(\frac{G_0(K+\frac{1}{2})-n}{\sqrt{n(1-\theta_0)}}) \stackrel{set}{=} \alpha$. Then we solve it and have $\frac{\theta_0(k+\frac{1}{2})-n}{\sqrt{n(1-\theta_0)}} \approx -z_\alpha$, which gives us $K \approx \frac{1}{\theta_0}[n-\sqrt{n(1-\theta_0)}z_\alpha]-\frac{1}{2}$.

7.2 U.M.P. Two-sided Tests

Go back to Table of Contents. Please click TOC

Consider first, $\mathcal{N}(\theta, \sigma_0^2)$, θ unknown, σ_0^2 known. We want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. There does not exist a U.M.P. level-α case. For $\theta < \theta_0$ we maximize the power at θ by rejecting H_0 for $\bar{x} < \theta_0 - \frac{\sigma_0}{z_\alpha}\sqrt{n}$, and for $\theta > \theta_0$, we maximize the power at θ by rejecting H_0 for $\bar{x} > \theta_0 + \frac{\sigma_0 z_\alpha}{\sqrt{n}}$. Since these are two different tests, we cannot maximize the power for both $\theta < \theta_0$ and $\theta > \theta_0$ by a single test.

In this situation, the standard approach is to break up α into 2 equal pieces. We reject H_0 if either $\bar{x} > \theta_0 + \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}$ or $\bar{x} < \theta_0 - \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}$. This test can be shown to be "U.M.P. unbiased". It maximizes the power at all θ among level$\leq \alpha$ tests which satisfy the additional requirement that $\pi_\theta = P_\theta(reject) \geq P_{\theta_0}(reject)$, for all $-\infty < \theta < \infty$.

We can also describe the above test by rejecting H_0 if $|z| = \sqrt{n}(\frac{|\bar{x}-\theta_0|}{\sigma_0}) > z_{\alpha/2}$. This test rejects H_0 at level-α if and if θ_0 does not belong to the two-sided confidence interval, $\bar{x} \pm \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}$.

More generally, if we have M.L.R. in a statistic T then for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, then w ereject H_0 for $T > c$, or $T < c_2$, where $P_\theta(T > c_1) = P_\theta(T < c_1) = \frac{\alpha}{2}$ in the discrete case we try to get close to $\alpha/2$ for both probabilities.

We already considered, $\mathcal{N}(\theta, \sigma_0^2)$ with σ_0^2 known. We will refer to that as example one.

7.3 χ^2 Goodness of Fit Tests

Go back to Table of Contents. Please click TOC

Let $X_1, ..., X_n$ be i.i.d., each X_i takes on values $1, ..., n$, with probabilities $p_1, ..., p_m$, $(p_j > 0, \sum_j p_j = 1)$. Let N_j denote the number of X_i which equal j, $j = 1, ..., m$. Frequently in applications the outcomes $1, ..., m$, represent categories, and need not be numerical. For example, red, blue, green, orange, which means that $1, ..., m$ are arbitrary labels and are not indicative of quantitative properties. In this case the data is known as categorical data.

We wish to test a model under which $P_i = p_i^*$, while $i = 1, .., m$. We want to see whether the outcomes are in line with the model.

We ewish to test $H_0 : p_i = p_i^*$ for $i = 1, ..., m$ versus $H_1 : p_i \neq p_i^*$ for some i. Under H_0, $\mathbb{E}(N_j) = \cap p_j^*$, $j = 1, ..., m$. The chi-square goodness of fit statistic is defined as

$$\chi^2 = \sum_{i=1}^{m} \frac{(N_i - np_i^*)^2}{np_i^*} = \sum_{i=1}^{n} \frac{(Observed - Example)_i^2}{Expected_i}$$

Under H_0 as $n \to \infty$, the distribution of χ^2 converges to χ^2_{m-1}. The chi-square goodness of fit test rejects H_0 at level α if $\chi 2 > \chi^2_{m-1,\alpha}$. If

the observed value of $\chi^2 = \chi$, then the approximate p-value is given by $p(\chi^2_{m-1} \geq \chi)$.

For the variations, often times we have a model with two parameters, θ_1, .., θ_q for which under H_0, we have $p_j = p_j(\theta_1 \dots \theta_q)$ with $q < m - 1$. The values of $\theta_1 \dots \theta_q$ are unspecified by the model. They need to be estimated from the data. Suppose that $\hat{\theta}_1, ..., \hat{\theta}_q$ are the M.L.E.'s of $\theta_1, ..., \theta_q$. Then $\mathbb{E}_{H_0}(N_j)$ is defined as $np_0(\hat{\theta}_1, \dots \hat{\theta}_j) = n\hat{p_j}$. We now use a modified χ^2 statistic,

$$\chi^2 = \sum_{j=1}^{m} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j}$$

Under H_0 as $n \to \infty$, $\chi^2 \to \chi^2_{m-1-q}$, a chi-square distribution with $m - 1 - q$ degrees of freedom. We lose one degree of freedom for each parameter estimated from the data.

Contingency Tables We cross classify observations according to two attributes. Factor A has levels $1, ..., a$, and Factor B has levels $1, ..., b$. Every one of the n observations is a pair (i, j), indicating the level (i) of Factor A and the level (j) of Factor B. Let N_w denote the number in the sample (of size n) classified as (i, j). We arrange the counts in a "contingency table". Let N_i be the total number for each row or column. Then we have

..	#	...	b	..
1	N_{11}	...	N_{1b}	N_{1j}
2	N_{21}	...	N_{2b}	N_{2j}
$\vdots$	...			
a	N_{a1}	...	N_{ab}	N_{aj}
..	N_{i1}	...	N_{ib}	n

The columns sum up each N_{ij} to be the total number classified in jth level of Factor B.

Let $P_{i,j}$ denote the true probability of an (i, j) classification. Under the model of independence between factors, $p_{ij} = \alpha_i \beta_j$, where α_i is the true probability of being classified at the ith level of A, and β_j the probability of the j th level of B. Since $\underset{i=1}{\overset{a}{\alpha}} \alpha_i = 1$, there are $a - 1$ free parameters, $\alpha_1, ..., \alpha_{a-1}$, similarly there are $b - 1$ free parameters $\beta_1, ..., \beta_{b-1}$. Thus, the total number of parameters in the independence model is $q = a+b-2$, and the chi-square statistic will be $m - 1 - q = ab - 1 - (a + b - 2) = ab - a - b + 1 = (a - 1)(b - 1)$.

Under H_0 the joint p.m.f. of the data equals

$$f(N_{ij} \dots |\theta) = c \prod_{i,j} (\alpha_i, \beta_j)^{N_{i,j}}$$

$$= C \left(\prod_{i=1}^{a-1} \alpha_i^{N_i} \right) \left(\prod_{j=1}^{b-1} \beta_j^{N_j} \right) \left(1 - \sum_{i=1}^{a-1} \alpha_i \right)^{N_{aj}} \left(1 - \sum_{j=1}^{b-1} \beta_j \right)^{N_{ib}}$$

That is, we have

$$\log f = \tilde{C} + \sum_{i}^{a-1} N_i \, log \, \alpha_i + N_a \log(1 - \sum_{i}^{a-1} \alpha_i) + \sum_{i}^{b-1} N_{ij} \log \, \beta_j + N_b \log(1 - \sum_{j}^{b-1} \beta_j)$$

then take derivative

$$\Rightarrow \frac{\partial}{\partial \alpha_i} \log f = \frac{N_i}{\alpha_i} - \frac{N_a}{\alpha_a} \Rightarrow \hat{\alpha}_i = \frac{N_i}{N_a} \hat{\alpha}_a$$

Thus, we have

$$1 - \hat{\alpha_1} = \sum_i^{n-1} \hat{\alpha}_n = \frac{\hat{\alpha}_a}{N_a} \sum_i^{a-1} N_i = \frac{\hat{\alpha}_a}{N_j}(n - N_a) = \frac{n\hat{\alpha}_a}{N_a} - \hat{\alpha}_a$$

which gives us

$$\Rightarrow \hat{\alpha}_a = \frac{N_1}{n}$$

and $\hat{\alpha}_i = \frac{1}{n}\hat{N}_i$ for $i = 1, ..., a-1$. Thus, the M.L.E. of $\hat{\alpha}_i$ is the sum of the elements in row i divided by n, which isthe observed proportion of observations in the ith level of Factor A.

Similarly, $\hat{\beta}_j = \frac{1}{n}N_j$ for $j = 1, .., b$ and then

$$\hat{p}_{ij} = \hat{\alpha}_i \hat{\beta}_j = \frac{1}{n^2} N_i N_j$$

and $\mathbb{E}_{H_0} N_{ij} = n\hat{p}_{ij} = \frac{1}{n}\hat{N}_i \hat{N}_j$. The chi-square statistic is then, finally,

$$\chi^2 = \sum \frac{(N_{ij} - \frac{1}{n} N_i N_j)^2}{\frac{N_i N_j}{n}}.$$

Thus, we reject null hypothesis at level α if $\chi^2 > \chi^2_{(a-1)(b-1),\alpha}$. If we reject independence, then we conclude that Factors A and B are dependent.

We discuss some more examples on χ^2 Tests.

Example 7.4. *For a 2×2 table, denote it with its marginals as*

1	*a*	*b*	*a+b*
2	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*a+b+c=d*

Then we have

$$\chi^2 = \left(a - \frac{(a+b)(a+c)}{n}\right)^2 \frac{n^3}{(a+b)(a+c)(c+d)(b+d)}$$

The above is proved by straightforward algebra.

7.4 Non-Parametric Tests

Go back to Table of Contents. Please click TOC

Non-parametric statistics refers to statistical inference without the assumption that the unknown distribution belongs to a parametric family. For an example, we have two independent samples, $X_1, ..., X_n$, $Y_1, ..., Y_m$ and we wish to test the null hypothesis that their true distributions, F for $X_1, ..., X_n$, G for $Y_1, ..., Y_m$, are identical. A one-sided alternative is that F is stochastically larger than G meaning,

$$\bar{F}(t) \geq \bar{G}(t), \forall t,$$

with inequality for some t. we assume that F and G are continuous. That avoids dealing tied observations. However, there are satisfactory methods for dealing with ties.

Thus, we have $H_0 : F = G$ versus $H_1 : F \geq G$ (while F stochastically greater than G), with F and G continuous. Note that we do not treat F and G as normally distributed and nor do we employ any other parametric family.

The following procedure was introduced by Wilcoxon in the early 1950's. To illustrate, suppose that $n = 6$, and $m = 5$. The observations are, for example, I for 65.2, 67.1, 69.4, 78.2, 74, 83; and II for 59.4, 72.1, 68, 66.4, 54.5. Rank the observations in the combined sample from smallest to largest giving ranks $1, ..., n+m$, in this case $1, ..., 11$ would be that group I has elements rank 3,5,7,10,9,11; and group II has elements rank would be 2,8,6,4,1.

Notice that the ranks in group I tend to be higher than in group II. If H_0 were true, then all $\binom{11}{6} = 462$ subsets of size 6 from $\{1, ..., 11\}$ would be equally likely as the set of ranks for the group I observations.

Define W, the Wilcoxon statistic to be the sum of the group 1 ranks; in this example $W = 3 + 5 + 7 + 10 + 9 + 11 = 45$. The idea is that if W is unusually large under H_0, then either it is due to change or it is due to H_0 being false in the direction of F being larger than G. In this example, the p-value corresponding to $W = 45$ equals, the number subsets of size 6 from $\{1, ..., 11\}$ with $W \geq 45$ divided 462. Then we compute

5th Rank	*6th Rank*	*Number of Subsets*	W
$11, 10, 9, 8, 7$	$1 - 6$	6	$46 - 51$
$11, 10, 9, 8, 6$	$1 - 5$	5	$45 - 49$
$11, 10, 9, 8, 5$	$2 - 4$	3	$45 - 47$
$11, 10, 9, 8, 4$	3	1	45
$11, 10, 9, 7, 6$	$2 - 5$	4	$45 - 48$
$11, 10, 9, 7, 5$	$3 - 4$	2	$45 - 46$
$11, 10, 9, 6, 5$	4	1	45
$11, 10, 8, 7, 6$	$3 - 5$	3	$45 - 47$
$11, 10, 8, 7, 5$	4	1	45
$11, 9, 8, 7, 6$	$4 - 5$	2	$45 - 46$
$11, 9, 8, 7, 6$	5	1	45

The p-value would be the total number of subsets of subsets of size 6 from $\{1, ..., 11\}$ with $W \geq 45$ divided by 462, which be $29/462 \approx= 0.628$.

We are thus led to doubt that H_0 is true, however we accept H_0 at the 0.05 level of significance, rejecting it for example at the 0.1 level.

For large values of n, m, enumeration is difficult. However, the normal approximation with continuity correction is qutie accurate.

Definition 7.5.

$$\delta_i = \begin{cases} 1 & , \quad if\ ith\ ranking\ observation\ from\ group\ I \\ 0 & , \quad if\ group\ II \end{cases}$$

Then, we define $W = \sum_{i=1}^{n+m} i\delta_i$, and we have

$$\begin{aligned}
\mathbb{E}_{H_0}(W) &= \sum_{i=1}^{n+m} i\mathbb{E}_{H_0}(\delta_i) \\
&= \sum_{i=1}^{n+m} i(\tfrac{n}{n+m}) \\
&= \tfrac{n}{n+m}\sum_{i=1}^{n+m} i \\
&= \tfrac{n}{n+m}\tfrac{(n+m)(n+m+1)}{2} \\
&= \tfrac{1}{2}n(n+m+1)
\end{aligned}$$

Next we have a property about variance.

Proposition 7.6. *We have* $var_{H_0}W = \sum_{i=1}^{n+m} i^2 var_{H_0}\delta_i + 2\sum^{1\le i<j\le n+m} ijcov(\delta_i,\delta_j)$.

Then we have

(1) $var_{H_0}W = (var_{H_0}\delta_i)\sum_{i=1}^{n+m} i^2 + 2cov(\delta_1,\delta_2)\sum^{1\le i<j\le n+m} ij$.

(2) Next, $var_{H_0}(\delta_i) = var_{H_0}(\delta_1) = \frac{n}{n+m}(1-\frac{n}{n+m}) = \frac{nm}{(n+m)^2}$ *while for* $i \neq j$, $cov_{H_0}(\delta_i,\delta_j) = cov(\delta_1,\delta_2) = \mathbb{E}_{H_0}(\delta_1,\delta_2) - (\mathbb{E}_{H_0}\delta_1)^2$.

(3) $var_{H_0}W = \frac{n}{n+m}\frac{n-1}{n+m-1} - \frac{n}{(n+m)^2} = -\frac{nm}{(n+m)^2(n+m-1)}$

(4) $\sum_{i=1}^{n+m} i^2 = \frac{1}{6}(n+m)(n+m+1)(2n+2m+1)$

(5) $2\sum_{1\le i<j\le n+m} ij = (\sum_{i=1}^{n+m} i)^2 - \sum_{i=1}^{n+m} i^2 = \frac{1}{12}(n+m)(n+m+1)[2(n+m)(n+m+1)-2(2n+2m+1)] = \frac{1}{12}(n+m(n+m+1)[3(n+m)+2][n+m-1]$

(6) Finally, we conclude

$$var_{H_0}W = \frac{1}{12}nm(n+m+1)$$

7.5 Paired T Test

Go back to Table of Contents. Please click TOC

In order to compare two population mean in the case of two samples that are correlated. We use paried sample t-test in "before-after" studies.

For the test, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ under some level of significance, say 5% (or sometimes say 1%). We calculate

$$t = \frac{\bar{d}}{\sqrt{s^2/n}}$$

, where $\bar{d}$ is the mean difference between two samples, that is, $\bar{d} = |\mu_1 - \mu_2|$. Notate s^2 to be the sample variance, n to be the sample size, and g to be the paried sample t-test with $n-1$ degrees of freedom. A formula for paired sample t-test is the following

$$t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2)-(\sum d)^2}{n-1}}}.$$

After the caclulation, we compare the t-value with table of critical values. If the calculuated value is greater than the table value, then we reject the null hypothesis for the paired sample t-test. If the calculated value is less than the table value, then we will accept the null hypothesis and say there is no significant mean difference between the two paired samples.

For two-sided tests, suppose we want to test $H_1 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$. Reject if $\frac{\sqrt{n}|\bar{d}|}{s_d} > t_{n-1,\alpha/2}$. This is equivalent as rejecting null hypothesis if either (i) $\bar{d} > \frac{s_d}{\sqrt{n}} t_{n-1,\alpha/2}$ or (ii) $\bar{d} < -\frac{s_d}{\sqrt{n}} t_{n-1,\alpha/2}$. This test is equivalent to saying rejecting H_0 at level α if 0 does not belong to the $100(1-\alpha)\%$ confidence interval, $(\bar{d} - \frac{s_d}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{d} + \frac{s_d}{\sqrt{n}} t_{n-1,\alpha/2})$.

Remark 7.7. For example, if we give training to a company employee and we want to know whether or not the training had any impact on the efficiency of the employee, we could use the paired sample test. We collect data from the employee on a seven scale rating, before the training and after the training. By using the paired sample t-test, we can statistically conclude whether or not training has improved the efficiency of the employee. In medicine, by using the paired sample t-test, we can figure out whether or not a particular medicine will cure the illness.

7.6 Proportional Hazard

Go back to Table of Contents. Please click TOC

Supposee that $\bar{F}_t(t) = P_e(X > t) = \bar{F}^{\theta}_{\theta=1}(t) = (\bar{F}(t))^{\theta}$, where $\bar{F}$ is the survival function of a continuous distribution on $(0, \infty)$. This is known as a proportional hazards family since

$$h_\theta(t) = \frac{d}{dt}(-\log \bar{F}_t(t)) = \frac{d}{dt}(-\theta \log \bar{F}(t))$$

and also we can write

$$h_\theta(t) = \theta \frac{f(t)}{\bar{F}(t)} = \theta h_{\theta=1}(t) = \theta h(t).$$

Thus, for eaech θ, h_θ the hazard function of F_θ is proportional to $h(t)$ the hazard function of F.

We are interested i ninference questions about θ from data $X_1, ..., X_n$ i.i.d. as $F(t)$ with θ unknown.

Definition 7.8. Define $H_\theta(t) = -\log \bar{F}_t(t) = \theta(-\log \bar{F}(t)) = \theta H(t)$, where H is the integrated hazard corresponding to F. Note that

$$\begin{aligned} P_\theta(H(X) > t) &= P_t(X > H^{-1}(t)) \\ &= e^{-\theta H(H^{-1}(t))} \\ &= e^{-et}, \ t > 0 \end{aligned}$$

We can thus transform $X_1, ..., X_n$ to $Y_1, ..., Y_n$ and do inference on exponential distributions.

Let us discuss an alternative way of describing the model. Let $\lambda(t|X_{1i}, X_{2i}, ..., X_{Ki})$ denote the hazard function for ith person at time t, $i = 1, ..., n$, where the

K regressors are denoted as $X_{1i}, X_{2i}, ..., X_{Ki}$. The baseline hazard function at time t, i.e., when $X_{1i} = 0$, $X_{2i} = 0$, ..., $X_{Ki} = 0$, is denoted as $\lambda_0(t)$. The baseline hazard function is analogous to the intercept term in a multiple regression or logistic regression model. Notice the baseline hazard function is not specified, but must be positive. The hazard ratio, $\lambda_1(t)/\lambda_0(t)$ can be regarded as the relative risk of the event occurring at time t, is a linear combination of parameters and regressors, i.e.,

$$log\left(\frac{\lambda(t|X_{1i}, X_{2i}, ..., X_{Ki})}{\lambda_0(t)}\right) = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}.$$

The ratio of hazard functions can be considered a ratio of risk functions, so the proportional hazards regression model can be considered as function of relative risk (while logistic regression models are a function of an odds ratio). Change in a coravriate have a multiplicative effect on the baseline risk. The model in terms of the hazard function at time t is:

$$\lambda(t|X_{1i}, X_{2i}, ..., X_{Ki}) = \lambda_0(t)exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_K X_{Ki}).$$

Although no particular probability model is selected to represent the survival times, proportional hazards regression does have an important assumption: the hazard for any individual is a fixed proportion of the hazard for any other individual (i.e., proportional hazards). Notice if $\lambda_0(t)$ is the hazard function for a subject with all the predictor values equal to zero and $\lambda_1(t)$ is the hazard function for a subject with other values for the predictor variables, then the hazard ratio depends only on the predictor variables and not on time t. This assumption means if a covariate doubles the risk of the event on day one, it also doubles the risk of the event on any other day.

7.7 Inequalities

Go back to Table of Contents. Please click TOC

First, we discuss markov's Inequality. For $x \geq 0$, $\mathbb{E}(X) = \mu < \infty$. Then for $a \geq \mu$, we have

$$P(X \geq a) \leq \frac{\mu}{a}$$

Notice that the bound is sharp. Consider the following definition

$$\mathbb{I} = \begin{cases} 1 & , \quad if\ x \geq a \\ 0 & , \quad if\ x < a \end{cases}$$

Then we have $0 \leq X \leq aI$, thus we have $0 \leq \mathbb{E}(X) = \mathbb{E}(a\mathbb{I}) = a\mathbb{P}(X \geq a)$.

There are few consequences:

(i) X are not necessarily positive. Assume that $\mathbb{E}|x| < \infty$. Then we have

$$\mathbb{P}(|X| \geq a) \leq \frac{1}{a}\mathbb{E}|x|$$

(ii) Consider $\mathbb{E}|x| < \infty$ and $\mu = \mathbb{E}(X)$. Note that $\mathbb{E}|x - \mu| =\leq \mathbb{E}|X| + |\mu| < \infty$. Applying (i) with X replaced by $X - \mu$. Then $\mathbb{P}(|X - \mu| \geq a) \leq \frac{\mathbb{E}|X-\mu|}{a}$, for $a \geq \mathbb{E}|X - \mu|$.

(iii) Suppose that $\mathbb{E}X^2 < \infty$, then $\sigma^2 = \mathbb{E}(X-\mu)^2 < \infty$. Use Markov with $(X-\mu)^2$. Then we have

$$\mathbb{P}(|X-\mu| \geq a) = \mathbb{P}((X-\mu)^2 \leq a^2) \leq \frac{\mathbb{E}(X-\mu)^2}{a^2} = \frac{\sigma^2}{a^2}$$

for $a \geq \sigma$. This is Chebychev's ineuqality. It is equivalent to $\mathbb{P}(a - \mu < X < a + \mu) \geq 1 - \frac{\sigma^2}{a^2}$ for $a \geq \sigma$.

(iv) $X_1, ..., X_n$ be i.i.d. with mean μ and variance σ^2. Let $\bar{X}_0$ denote the sample mean. Apply (iii), with $\mathbb{E}(\bar{X})_n = \mu$, $Var(\bar{X}_n) = \sigma^2/n$. Then $\mathbb{P}(a - \mu < \bar{X}_n < a + \mu) \geq 1 - \frac{\sigma^2}{na^2} \to 1$ as $n \to \infty$. Or equivalently, for every $\epsilon > 0$, there is $\lim_{n\to\infty} Pr(|\bar{X}_n - \mu| \geq \epsilon) = 0$. This result is the "weak law of large numbers".

8 Regression

Go back to Table of Contents. Please click TOC

Consider data of the form

$$\binom{x_1}{y_1} \cdots \binom{x_n}{y_n}$$

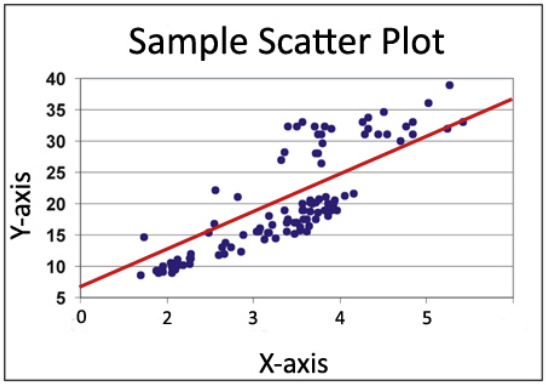
The vectors are considered independent, however within each vector x_i and y_i may be dependent. It is believed that an anverage x and y are linearly related, but not in a deterministic sense. The model, the mean of the conditional distribution of Y given $X = x$, is $\mathbb{E}(Y|X = x) = \alpha + \beta x$. The variance $Var(Y|X = x) = \sigma^2$, independent of x is given. These assumptions are summarized by

$$\mathbb{E}(Y|X = x) = \alpha + \beta x + \epsilon$$

with ϵ the "error" or "noise", having mean zero and variance σ^2. Later we will also assume that ϵ is normally distributed.

The variables Y and X are not treated symmetrically. The X variable is referred to as an explanatory or independent variable, while the Y variable is called a dependent or response variable. In most of the work, the X's are treated as observed random variables, $X_i = x_i$ for $i = 1, 2, ..., n$, and then $Y_1, ..., Y_n$ are studied in the context of their conditional behavior given the x_n's. A scatter plot, plots all the observed pairs, $\binom{x_i}{y_i}$, for $i = 1, 2, 3, ..., n$, can be some random plot on the cartesian coordinates.

Figure 2: Graphic illustration for a random plot that fits for a linear model.



From the sample plot in the graph above, we get the senses that as x increases y tends to increase as well. Other scatter plots may convey different relationships and magnitude of relationship between x and y.

8.1 Statistics in Regression

Go back to Table of Contents. Please click TOC

Based on the understanding of regression, we can introduce some statistics. Consider random sample $x_1, x_2, ..., x_n$, we have mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and variance

$$s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We consider $s_x^2/(n-1)$ to be an unbiased estimator of $Var(x)$. Similarly, we can consider $y_1, y_2, ..., y_n$ and we have $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and $s_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$. We also use that

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

with $\frac{s_{xy}}{n-1}$ an unbiased estimator of $Cov(X,Y) = \mathbb{E}(X - \mathbb{E}(x))(Y - \mathbb{E}(Y))$. It does not matter if we divide by n or $n-1$ as long as we are consistent. The statistic,

$$\hat{p} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}},$$

is an estimator of the correlation coefficient between X and Y.

8.2 Fitting a Straight Line to the Data

Go back to Table of Contents. Please click TOC

For a given α, β, define

$$C(\alpha,\beta)=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2,$$

the squared Euclidean distance between the points $\binom{x_1}{y_1}\dots\binom{x_n}{y_n}$ and the potential straight line predicts $\binom{x_1}{\alpha+\beta x_i}\dots\binom{x_n}{\alpha+\beta x_n}$. We seek the values of α and β (call them "optimal" values A and B). Then $C(A,B)<C(\alpha,\beta)$ for all $(\alpha,\beta)\neq(A,B)$. The resulting line, $\hat{y}(x)=A+Bx$, is known as the least squares fitted straight line, with B the fitted slope and A the fitted y intercept. It turns out $B=\frac{s_{xy}}{s_x^2}$ and $A=\bar{Y}-\frac{s-xy}{s_x^2}\bar{x}$. The least squares straight line is given by

$$\hat{y}(x)=A+Bx=\bar{y}+\frac{s_{xy}}{s_x^2}(s_i-\bar{x})$$

Example 8.1. *Consider the following example with three observations. Suppose we observe* $\binom{x_1}{y_1}=\binom{1}{5}$, $\binom{x_2}{y_2}=\binom{2}{6}$, *and* $\binom{x_3}{y_3}=\binom{3}{6}$. *We can compute* $\bar{x}=(1+2+3)/3=2$, $\bar{y}=(5+6+16)/3=9$. *This gives us* $s_x^2=2$ *and* $s_y^2=74$. *Thus, we have* $B=\frac{s_{xy}}{s_x^2}=11/2=5.5$ *and* $A=\bar{y}-\frac{s_{xy}}{s_x^2}=9-11=-2$.

Thus, we have least squares fitted line:

$$\hat{y}(x)=-2+5.5x.$$

Next, let us take a step back to look at our data: $y_1=5$, $y_2=6$, $y_3=16$. *We can use least squares fitted line to approximate these values, which would be* $\hat{y}_1=3.5$, $\hat{y}_2=9$, *and* $\hat{y}_3=13.5$ *with* $\sum_{i=1}^{3}(\hat{y}-y_i)^2=2.25+9+2.25=13.5$.

Remark 8.2. Notice that in the absence of the x's, the least squares fitting horizontal line is $\hat{y}=\bar{y}$. That is because $\min_c\sum(y_i-c)^2=\sum(y_i-\bar{y})^2=s_y^2$. Thus, by observing x and using the best fitting straight line, the original uncertainty in Y, $s_y^2=74$, is now reduced to $\sum_{i=1}^{3}(\hat{y}_i-y_i)^2=13.5$.

Remark 8.3. A useful identity to remember is

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-y_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

References

[1] Ross, Sheldon M., *Introduction to Probability and Statistics for Engineers and Scientists.*