STAT230

PROBABILITY

Chapter 9 (b)

UNIVERSITY OF WATERLOO

Relationships Between Variables

- Independence is a “yes/no” way of defining a relationship between variables.

- There can be different types of relationships between variables which are dependent.

  - Deterministic Relationship.

  - Statistical Relationship.

Deterministic Relationship vs. Statistical Relationship

➢ Deterministic relationship has no variance nor uncertainty.

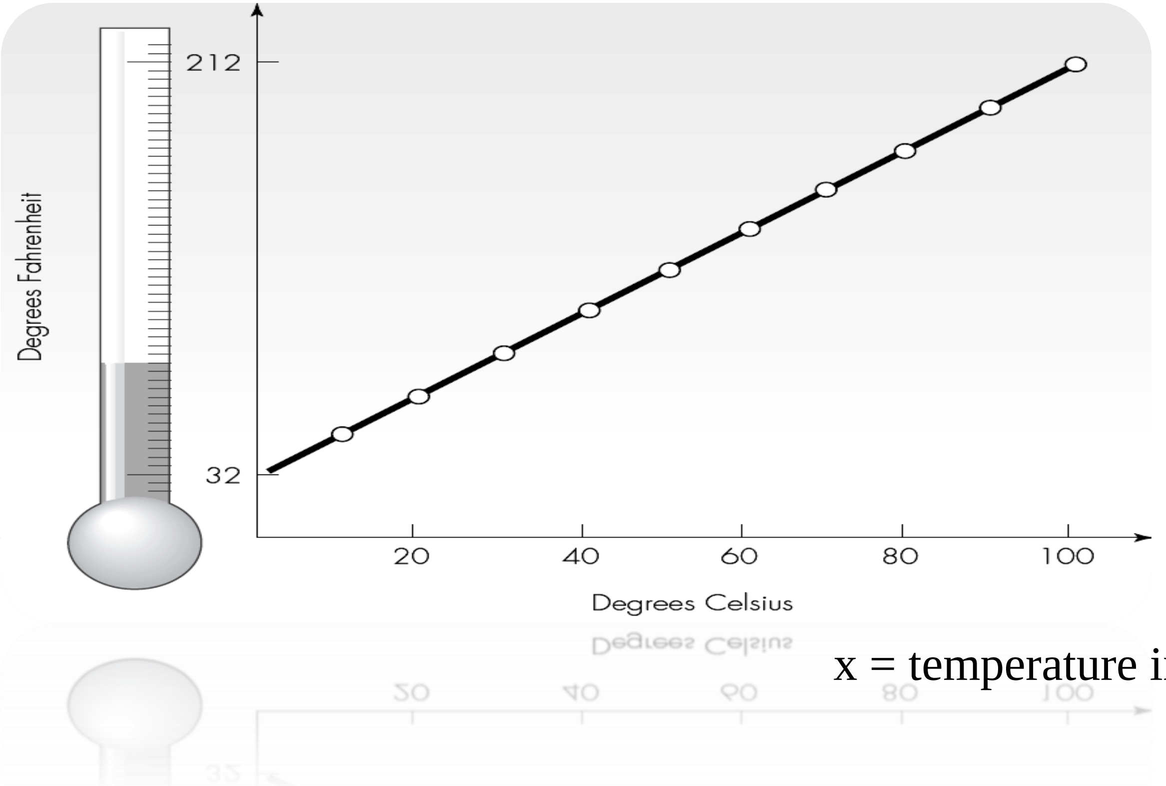
- If we know the value of one variable, we can determine the value of the other *exactly*.
  - Unit conversions (E.g., miles to kilometers, meter to cm)
  - Known scientific formulas (E.g. water volume and weight)

- The simplest deterministic mathematical relationship between two variables x and y is a linear relationship.

$$y = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} x$$

Deterministic Relationship

y = temperature in Fahrenheit



x = temperature in Celsius

$$y = 32 + 1.8x$$

- Statistical relationship(**Probabilistic** Non Deterministic)

- Has both variance and uncertainty(***randomness)*** that is part of a real-life process.
- For a fixed value of x, there is uncertainty in the value of the second variable y.

Example

x = high school grade point average (GPA).

y = collage GPA.

Letter Grade	Grade Points
A+	4.0
A	4.0
A-	3.7
B+	3.3
B	3.0
B-	2.7
C+	2.3
C	2.0
C-	1.7
D+	1.3
D	1.0
F	0.0

- We'll look at two ways of measuring the strength of the relationship between two random variables:

  - The covariance.

  - The correlation coefficient.

Covariance

- The *covariance* between two r.v's X and Y is

E(X)

$$\operatorname{Cov}(X,Y)=E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$=\begin{cases} =\sum_x \sum_y (x-\mu_X)(y-\mu_Y)f(x,y) & \text{discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy & \text{continuous} \end{cases}$$

$$\operatorname{Cov}(X,Y)=E(XY)-\mu_X\cdot\mu_Y$$

UNIVERSITY OF WATERLOO

The *covariance* is a measure of how much two r.v.s change together.

- **Cov(X,Y) > 0**

If the **large values** of one variable tend to occur with **large values** of the other variable, and the same holds for the smaller values, (*the variables tend to show similar behavior*), the covariance is positive (positive relationship).

UNIVERSITY OF WATERLOO

- **Cov(X, Y) < 0**

If **small values** values of one variable tend to occur with **large values** of the other, (the variables tend to show opposite behavior), the covariance is negative (negative relationship) .

Example

X = number of cigarettes smoked/day.

Y = length of life.

- If X and Y are not strongly related, positive and negative products will tend to cancel one another, yielding a covariance near 0

Proposition

- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y, X)$

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

- $\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X,Y)$

- $\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$

- $Cov(X+Y, X-Y) = Var(X) - Cov(X,Y) + Cov(Y, X) - Var(Y)$

  $= Var(X) - Var(Y)$

- $Cov(aX + bY, cU + dV) =$

$acCov(X, U) + adCov(X, V) + bcCov(Y, U) + bdCov(Y, V)$

UNIVERSITY OF WATERLOO

Note

If X and Y are independent, then Cov(X,Y) = 0. However, the converse is not true.

Example

Suppose $P(X = 0) = P(X = 1) = P(X = -1) = 1/3$

and defining

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{if } X = 0 \end{cases}$$

Now, $XY = 0$, so $E(XY) = 0$. Also, $E(X) = 0$.

Thus, $Cov(X,Y) = E(XY) - E(X)\,E(Y) = 0$

However, X and Y are clearly not independent.

UNIVERSITY OF WATERLOO

Example

The joint and marginal p.f’s for X= automobile police deductible amount and Y=homeowner policy deductible amount find Cov(X,Y).

x\ y	0	100	200	Sum
100	0.2	0.1	0.2	0.5
250	0.05	0.15	0.3	0.5
Sum	0.25	0.25	0.5	1

UNIVERSITY OF WATERLOO

$$\mu_x = \sum_{\text{all } x} x\, f_X(x) = 175 \text{ all x}$$

$$\mu_y = \sum_{\text{all } y} y\, f_Y(y) = 125$$

$$\text{Cov}(X,Y) = \sum_{\text{all } x} \sum_{\text{all } y} (x - 175)(y-125)\, f(x,y)$$

$$= (100-175)(0-125)(0.2) + (100-175)(100-25)(0.1) + \ldots$$

$$\ldots. + (250-175)(100-125)(0.3)$$

$$=1875$$

The positive covariance means X tends to be positive when Y is positive and negative when Y is negative

Notes

- The magnitude of the covariance is not easy to interpret.
- The covariance computed value depends critically on the units of measurement.

  (meter = 100 centimeters, has higher Cov values than cm)

  But there wouldn't really be any increase in the degree to which X and Y are correlated; the change of units would just have spread things out.

- Ideally the choice of units should have no effect on a measure of strength of relationship.
- Solution: Scale the covariance

- Correlation Coefficient

- The *correlation coefficient* of X and Y, denoted by Corr(X, Y), $\rho_{X,Y}$ or just ρ is defined by

Note: $$\rho = \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

- For any two r.v.'s X and Y,

$$-1 \le \mathrm{Corr}(X,Y) \le 1$$

- The relationship will be described as

Strong if $|\rho| \geq 0.8$ **Moderate if $0.5 < |\rho| < 0.8$**

Weak $|\rho| \le 0.5$

UNIVERSITY OF WATERLOO

Correlation Proposition

- If X and Y are independent, then ρ =0 but ρ =0 does not imply independence.

  - When $\boldsymbol{\rho = 0}$ **,** X and Y said to be **uncorrelated.**

- For some numbers a and b with ρ =1 or -1 iff Y = aX + b where $a \neq 0$

- ρ is a measure of the degree of linear relationship between X and Y.

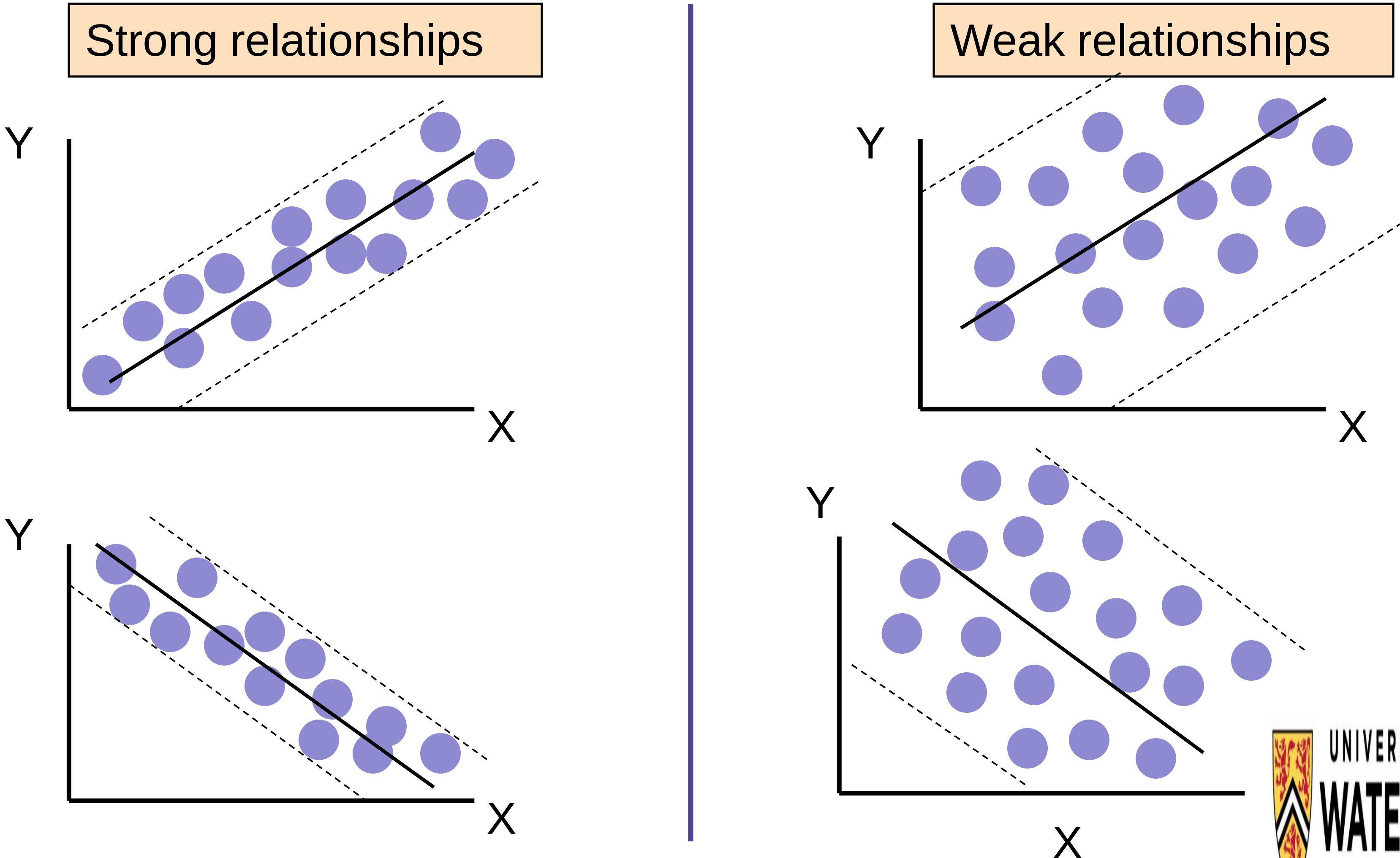
Note

- A value of   near 1 does not necessarily imply that increasing the value of X causes Y to increase.

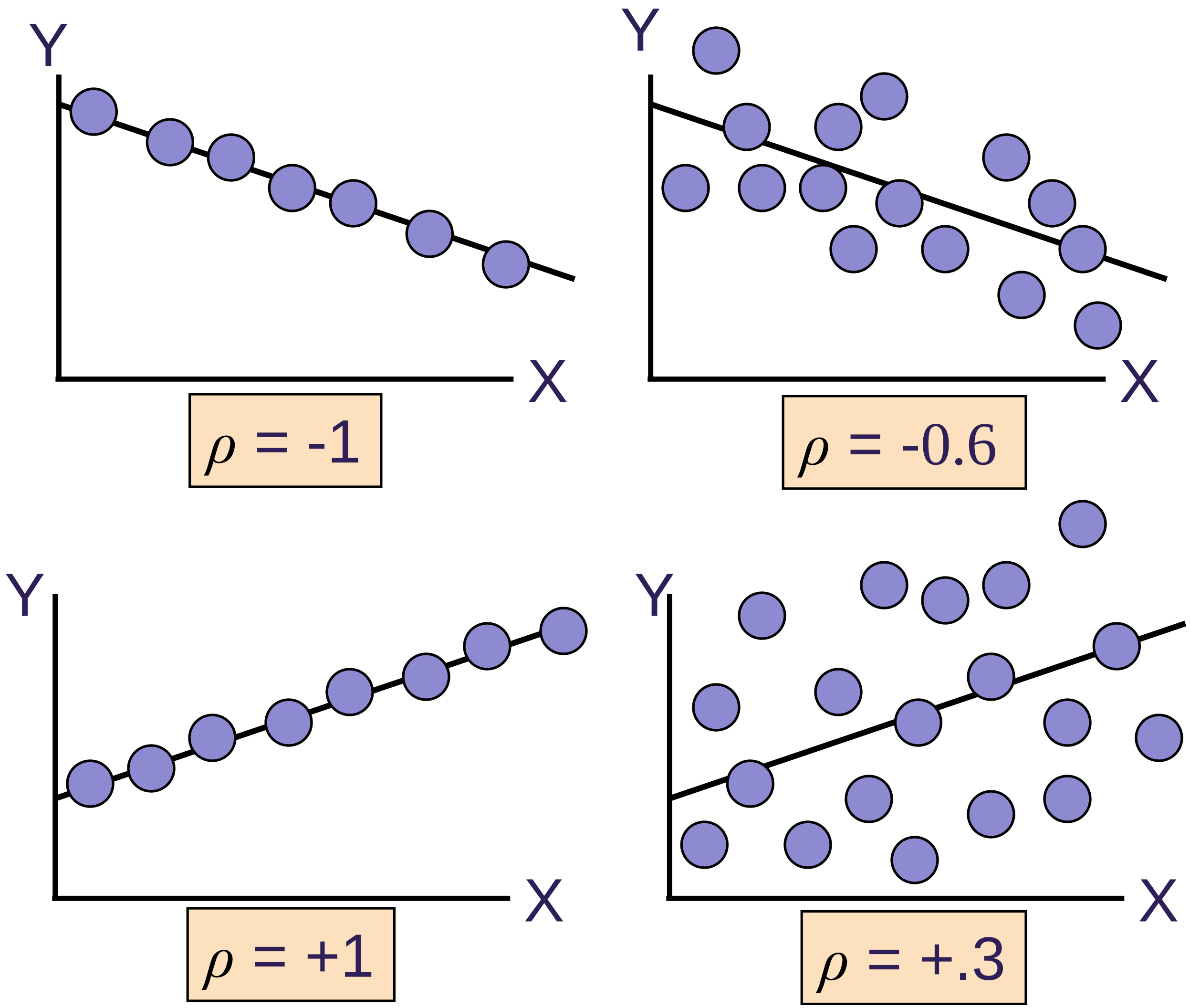
- A value of near 1 implies that large X values are associated with large Y values .

- Association is not the same as causation.

Linear Correlation



Scatter Plots of Data with Various Correlation Coefficients



UNIVERSITY OF WATERLOO

Example

Suppose Var(X) = 1.69, Var(Y) = 4, ρ= 0.5. Find the standard deviation of U = 2X – Y.

$$\rho = \rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = 0.5$$

Cov(X,Y)= 0.5 $\sqrt{1.96}\ \sqrt{4} =$ 1.3

Var(U) = V(2X – Y)= 4Var(X) +Var(Y) - 2(2) Cov(X,Y)

= 5.56

SD(U) = $\sqrt{5.56}$

Linear Combination

- Given a collection of n random variables $X_1,\ldots, X_n$ and n numerical constants (real numbers) $a_1,\ldots,a_n$, the r.v.

$$Y = a_1 X_1 + \ldots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

is called a linear combination of the X_i's.

Mean of a Linear Combination of Random Variables

Let $X_1,\ldots, X_n$ have mean values $\mu_1, \mu_2,\ldots,\mu_n$ and variances of $\sigma_1^2, \sigma_2^2, \ldots,\sigma_n^2$ respectively

Whether or not the X_i's are independent,

$$E\left(a_1X_1+\ldots+a_nX_n\right)=a_1E\left(X_1\right)+\ldots+a_nE\left(X_n\right)$$

$$=a_1\mu_1+\ldots+a_n\mu_n$$

UNIVERSITY OF WATERLOO

- Let a_i be constants (real numbers) and $E(X_i) = \mu_i$, $i = 1, 2, \ldots, n$. Then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mu_i$$

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

- Let $X_1, \ldots, X_n$ be random variables which have mean μ.

Let the sample mean $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then

$$E(\overline{X}) = \mu$$

UNIVERSITY OF WATERLOO

Variance of a Linear Combination of Random Variables

If X, Y are independent, then Cov(X,Y)=0

Var(X+Y)= Var(X) + Var(Y)

$$= \sigma^2_X + \sigma^2_Y$$

If $X_1,\ldots, X_n$ are independent,

$$V(a_1X_1 + \ldots + a_nX_n) = a_1^2V(X_1) + \ldots + a_n^2V(X_n)$$
$$= a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a^2_i \sigma^2_i$$

UNIVERSITY OF WATERLOO

If X, Y are independent, then $Cov(X,Y)=0$

$$Var(X - Y) = Var(X) + (-1)^2 Var(Y)$$
$$= \sigma^2_X + \sigma^2_Y$$

The variance of a difference is the sum of the variances.

If *X* and *Y* are dependent,

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov((X,Y)$$

Let a and b are constants then,

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov((X,Y)$$

If $X_1,\ldots, X_n$ are dependent,

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a^2_i \sigma^2_i + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} a_ia_j Cov(X_i, X_j)$$

Difference Between Two Random Variables

$$E(X_1 - X_2) = E(X_1) - E(X_2)$$

and, if X_1 and X_2 are independent with the same variance σ^2,

$$V(X_1 - X_2) = V(X_1) + V(X_2)$$

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i / n\right) = \sigma^2/n$$

UNIVERSITY OF WATERLOO

Notes

- The average $\bar{X}$ of n random variables with the same distribution is less variable than any single observation X_i, and that the larger n is the less variability there is.

- As $n \to \infty$; $Var(\bar{X}) \to 0$, which means that $\bar{X}$ becomes arbitrarily close to μ. This is called " The Law of Large numbers"

UNIVERSITY OF WATERLOO

Linear Combinations of Independent Normal Random Variables

- Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, where a and b are constant real numbers. Then

$$Y \sim N(a\mu + b, a^2\sigma^2).$$

- Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ independently, and let a and b be constants. Then

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$$

UNIVERSITY OF WATERLOO

- If $X_1,\ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

$$\overline{X} \sim N(\mu, \sigma^2 / n)$$

Example

Suppose $X \sim N(3, 5)$ and $Y \sim N(6, 14)$ independently. Find $P(X > Y)$.

Solution

$$P(X > Y) = P(X-Y > 0)$$

$$X - Y \sim N(3 - 6, 5 + 14) = N(-3, 19)$$

UNIVERSITY OF WATERLOO

$$P(X-Y > 0) = P\left(Z > \frac{0-(-3)}{\sqrt{19}}\right) \quad \text{where } Z \sim N(0,1)$$

$$= P(Z > 0.69)$$

$$= 1 - P(Z < 0.69)$$

$$= 1 - 0.75490$$

$$= 0.2451$$

UNIVERSITY OF WATERLOO

This table gives the values of $F(x)$ for $x \geq 0$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56750	0.57142	0.57534
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

Example

The heights of adult females in a large population is well represented by a Normal distribution with mean 64 inches and variance 6.2 (inches)^2.

- Suppose 10 women are randomly selected, and let $\overline{X}$ be their average height. Find $P(63 \leq \overline{X} \leq 65)$.

$\overline{X} \sim N(64, 6.2/10)$

$$P(63 \le \overline{X} \le 65) = P\left(\frac{63 - 64}{\sqrt{6.2/10}} \le Z \le \frac{65-64}{\sqrt{6.2/10}} \right)$$

$$= P(-1.27 \le Z \le 1.27) \text{ where } Z \sim N(0,1)$$

$$= 2P(Z \le 1.27) - 1$$

$$= 2(0.89796) - 1$$

$$= 0.79592$$

This table gives the values of $F(x)$ for $x \geq 0$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56750	0.57142	0.57534
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670

(b) How large must n be so that a random sample of n women gives an average height $\overline{X}$ so that $P(|\overline{X} - 64| \leq 1) \geq 0.95$?

Solution

$\overline{X} \sim N(64, 6.2/n)$

$$P(|\overline{X} - 64| \leq 1) = P\left(|Z| \leq \frac{1}{\sqrt{6.2/n}}\right)$$

$$= P\left(|Z| \leq \frac{\sqrt{n}}{\sqrt{6.2}}\right) \geq 0.95$$

$2P(Z \leq \sqrt{n} / \sqrt{6.2}) - 1 \geq 0.95$

$P(Z \leq \sqrt{n} / \sqrt{6.2}) \geq (1+0.95)/2 = 0.975$

UNIVERSITY OF WATERLOO

$P(|Z| \leq 1.96) \geq 0.95$

so, $\sqrt{n / 6.2} \geq 1.96$

$n \geq [1.96\,(6.2)\,]^2 = 23.77$ so $n \geq 24$

Example

Suppose that n independent measurements $X_1, X_2, \ldots, X_n$ are to be made to determine my height μ, an unknown quantity.

The mean of the measurements is to be used to estimate μ.

Suppose the errors in the measurement procedure are such that the X_i's are normally distributed with mean μ and standard deviation equal to 1 cm.

How large should n be to ensure, with probability 0.95, that the mean of the measurements is within 0.1 cm of μ?

UNIVERSITY OF WATERLOO

Since $X_i \sim N(\mu,(1)^2)$ then $\overline{X} \sim N\left(\mu,\frac{1}{n}\right)$. We want $P\left(\left|\overline{X}-\mu\right| \le 0.1\right) \ge 0.95.$

Therefore $0.95 \le P\left(\frac{\left|\overline{X}-\mu\right|}{1/\sqrt{n}} \le \frac{0.1}{1/\sqrt{n}}\right)$

$$= P\left(|Z| \le 0.1\sqrt{n}\right) \qquad \text{where } Z \sim N(0,1)$$

$$= 2P\left(Z \le 0.1\sqrt{n}\right) - 1$$

or $$P\left(Z \le 0.1\sqrt{n}\right) \ge \frac{(1+0.95)}{2} = 0.975.$$

From tables $P(Z \le 1.96) = 0.975$ so we want n such that

$$0.1\sqrt{n} \ge 1.96$$

or $$n \ge \left(\frac{1.96}{0.1}\right)^2 = 384.16.$$

Therefore at least 385 measurements should be taken.

This table gives the values of $F(x)$ for $x \geq 0$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56750	0.57142	0.57534
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670