Joint Probability Distributions (discrete case only)

The **joint probability mass function** of two discrete random quantities X, Y is

$$p(x, y) = \mathrm{P}[\, X = x \quad \text{AND} \quad Y = y \,]$$

The **marginal probability mass functions** are

$$\boxed{p_X(x) = \sum_y p(x, y)} = \mathbf{P}[X = x] \qquad \boxed{p_Y(y) = \sum_x p(x, y)}$$

Example 9.01

Find the marginal *p.m.f.*s for the following joint *p.m.f.*

$p(x, y)$	$y = 3$	$y = 4$	$y = 5$	$p_X(x)$
$x = 0$	.30	.10	.20	**.60**
$x = 1$	.20	.05	.15	**.40**
$p_Y(y)$	**.50**	**.15**	**.35**	**1**

[Check that both marginal *p.m.f.*s, (row and column totals), each add up to 1.]

The random quantities X and Y are **independent** if and only if

$$p(x, y) = p_X(x) \bullet p_Y(y) \quad \forall \; (x, y)$$

In example 9.01, $p_X(0) \bullet p_Y(4) = .60 \times .15 = .09$, but $p(0, 4) = .10$.
Therefore X and Y are **dependent**,
[despite $p(x, 3) = p_X(x) \bullet p_Y(3)$ for $x = 0$ and $x = 1$!].

For any two possible events A and B, conditional probability is defined by

$$\mathrm{P}[A \mid B] = \frac{\mathrm{P}[A \cap B]}{\mathrm{P}[B]}$$, which leads to the conditional probability mass functions

$$p_{Y|X}(y \mid x) = \frac{p(x, y)}{p_X(x)} \quad \text{and} \quad p_{X|Y}(x \mid y) = \frac{p(x, y)}{p_Y(y)} .$$

In example 9.01,

"$p_{Y|X}(5|0)$" means "P[$Y = 5 \mid X = 0$]".
"$p(0, 5)$" means "P[$X = 0$ and $Y = 5$]"

$$p_{Y|X}(5 \mid 0) = \frac{p(0,5)}{p_X(0)} = \frac{.20}{.60} = \frac{1}{3}$$

"$p_X(0)$" means "P[$X = 0$]"

Compare with P[$Y = 5$] = .35 :
events "$X = 0$", "$Y = 5$" are not quite independent!

$p(x, y)$	$y = 3$	$y = 4$	$y = 5$	$p_X(x)$
$x = 0$	.30	.10	**.20**	**.60**
$x = 1$	.20	.05	.15	.40
$p_Y(y)$	.50	.15	.35	1

Expected Value

$$\mathrm{E}[h(X,Y)] = \sum_x \sum_y h(x,y) \bullet p(x,y)$$

A measure of dependence is the **covariance** of X and Y:

$$\mathrm{Cov}[X,Y] = \mathrm{E}\big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\big] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x,y)$$

$$= \mathrm{E}[\,X\,Y\,] - \mathrm{E}[\,X\,] \bullet \mathrm{E}[\,Y\,]$$

Note that $\mathrm{V}[\,X\,] = \mathrm{Cov}[\,X, X\,]$.

In Example 9.01:

$p(x, y)$	$y = 3$	$y = 4$	$y = 5$	$p_X(x)$
$x = 0$	.30	.10	.20	.60
$x = 1$	.20	.05	.15	.40
$p_Y(y)$	.50	.15	.35	1

$$\mathrm{E}[X] = \sum_{x=0}^{1} x \cdot p_X(x) = 0\times.60 + 1\times.40 = \underline{\mathbf{0.40}}$$

$$\begin{aligned}\mathrm{E}[Y] &= \sum_{y=3}^{5} y \cdot p_Y(y) \\ &= 3\times.50 + 4\times.15 + 5\times.35 = \underline{\mathbf{3.85}}\end{aligned}$$

$$\mathrm{E}[XY] = \sum_{x=0}^{1}\sum_{y=3}^{5} xy \cdot p(x,y)$$

$x \times y$	**3**	**4**	**5**
0	$0 \times 3 \times .30$	$0 \times 4 \times .10$	$0 \times 5 \times .20$
1	$1 \times 3 \times .20$	$1 \times 4 \times .05$	$1 \times 5 \times .15$

$$\begin{aligned}&= \mathbf{0 + 0 + 0 + .60 + .20 + .75} \\ &= \underline{\mathbf{1.55}}\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathrm{E}[XY] - \mathrm{E}[X] \bullet \mathrm{E}[Y] \\ &= \mathbf{1.55 - 0.40\times3.85} \\ &= \underline{\mathbf{0.01}}\end{aligned}$$

Note that the covariance depends on the units of measurement. If X is re-scaled by a factor c and Y by a factor k, then

$$\begin{aligned}\mathrm{Cov}[cX, kY] &= \mathrm{E}[cXkY] - \mathrm{E}[cX] \bullet \mathrm{E}[kY] = ck\,\mathrm{E}[XY] - c\,\mathrm{E}[X] \bullet k\,\mathrm{E}[Y] \\ &= ck\,(\mathrm{E}[XY] - \mathrm{E}[X] \bullet \mathrm{E}[Y]) = ck \bullet \mathrm{Cov}[X, Y]\end{aligned}$$

A special case is $\mathrm{V}[cX] = \mathrm{Cov}[cX, cX] = c^2 \bullet \mathrm{V}[X]$.

This dependence on the units of measurement of the random quantities can be eliminated by dividing the covariance by the geometric mean of the variances of the two random quantities:

The **correlation coefficient** of X and Y is $\text{Corr}(X, Y) = \rho_{X,Y} =$

$$\rho = \frac{\text{Cov}[X,Y]}{\sqrt{\text{V}[X] \bullet \text{V}[Y]}} = \frac{\text{Cov}[X,Y]}{\sigma_X \bullet \sigma_Y}$$

In Example 9.01,

$$\text{E}\left[X^2\right] = \sum_{x=0}^{1} x^2 \cdot p_X(x) = 0^2 \times .60 + 1^2 \times .40 = \underline{\mathbf{0.40}}$$

$$\text{V}[X] = \mathbf{E[X^2] - (E[X])^2 = 0.40 - (0.40)^2 = 0.24}$$

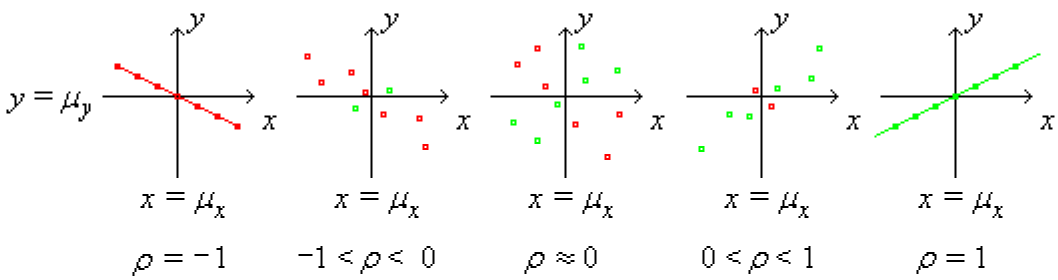
$$\text{E}\left[Y^2\right] = \sum_{y=3}^{5} y^2 \cdot p_Y(y) = \ldots = \underline{\mathbf{15.65}}$$

$$\text{V}[Y] = \mathbf{15.65 - (3.85)^2 = 0.8275}$$

$$\Rightarrow \quad \rho = \frac{0.01}{\sqrt{0.24 \times 0.8275}} \approx 0.0224$$

[See the Excel file "`www.engr.mun.ca/~ggeorge/3423/demos/jointpmf.xls`"].

For a joint uniform probability distribution: ***n*** **possible points,** $p(x, y) = 1/n$ **for each.**
(and noting $\rho \propto \sum_x \sum_y (x - \mu_x)(y - \mu_y)$):



[When $\rho = \pm 1$ exactly, then $Y = aX + b$ exactly, with sign(a) = ρ.]
In general, for constants a, b, c, d, with a and c both positive or both negative,

$$\text{Corr}(aX + b, \ cY + d) = \text{Corr}(X, \ Y)$$

Also: $\qquad -1 \le \rho \le +1.$

Rule of thumb:

$|\rho| \geq .8 \quad \Rightarrow \quad$ strong correlation

$.5 < |\rho| < .8 \quad \Rightarrow \quad$ moderate correlation

$|\rho| \leq .5 \quad \Rightarrow \quad$ weak correlation

In the example above, $\rho = 0.0224 \Rightarrow$ very weak correlation (almost uncorrelated).

X, Y are independent $\Rightarrow$ $p(x, y) = \mathbf{p_X(x)\, p_Y(y)}$

$\Rightarrow$ $\mathrm{E}[XY] = \sum\sum x\, y\, p(x)\, p(y) = \sum x\, p(x) \sum y\, p(y) = \mathrm{E}[X]\, \mathrm{E}[Y]$
$\Rightarrow$ $\mathrm{Cov}[X, Y] = \mathbf{E[XY] - E[X]\, E[Y]} = 0$. It then follows that
X, Y are independent $\Rightarrow$ X, Y are uncorrelated $(\rho = 0)$, but

X, Y are uncorrelated $\not\Rightarrow$ X, Y are independent .

Counterexample (9.02):

Let the points shown be equally likely. Then the value of Y is completely determined by the value of X. The two random quantities are thus highly dependent. Yet they are uncorrelated!

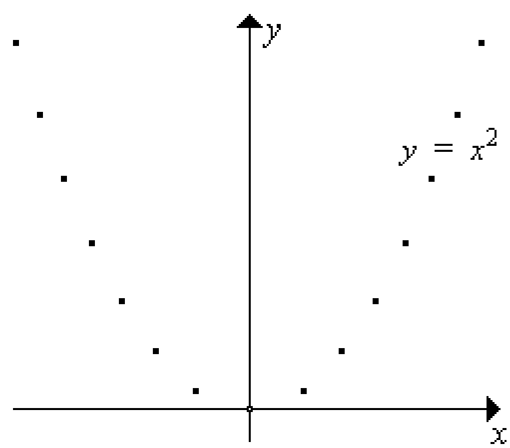
$\mathbf{Y = X^2}$ **(dependent)**

$\mathbf{E[X] = 0}$ **(by symmetry)**

$\mathbf{E[XY] = E[X^3] = 0}$ **(sym.)**

$$\mathbf{Cov(X, Y) = E[XY] - E[X]\, E[Y] = 0 - 0 \times E[Y] = 0}$$

Therefore $\boldsymbol{\rho = 0}$ **!**



[The line of best fit through these points is horizontal.]

Linear Combinations of Random Quantities

Let the random quantity Y be a linear combination of n random quantities X_i :

$$Y = \sum_{i=1}^{n} a_i X_i$$

[Note: lower case a_i is constant; upper case X_i is random]

then $\mathrm{E}[Y] = \mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathrm{E}[X_i]$ **(linear function)**

$$\therefore \quad \mu_Y = \sum_{i=1}^{n} a_i \mu_i$$

But $\mathrm{V}[Y] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{Cov}[X_i, X_j]$

$\{X_i\}$ independent $\Rightarrow$ $\underline{\mathrm{V}[Y] = \sum_{i=1}^{n} a_i^2 \mathrm{V}[X_i]}$

Special case: $n = 2$, $a_1 = 1$, $a_2 = \pm 1$:

$\mathrm{E}[X_1 \pm X_2] = \boldsymbol{\mu_1 \pm \mu_2}$ and

$\mathrm{V}[X_1 \pm X_2] = \mathbf{1^2\sigma_1^2 + (\pm 1)^2\sigma_2^2 = \sigma_1^2 + \sigma_2^2}$

["The variance of a difference is the <u>sum</u> of the variances"]

<u>Example 9.03</u>

Two runners' times in a race are independent random quantities T_1 and T_2 , with

$\mu_1 = 40$ $\sigma_1^2 = 4$,

$\mu_2 = 42$ $\sigma_2^2 = 4$,

Find $\mathrm{E}[T_1 - T_2]$ and $\mathrm{V}[T_1 - T_2]$.

$\mathbf{E}[\boldsymbol{T_1 - T_2}] = \boldsymbol{\mu_1 - \mu_2} = \underline{\underline{\mathbf{-2}}}$

$\mathbf{V}[\boldsymbol{T_1 - T_2}] = \boldsymbol{\sigma_1^2 + \sigma_2^2} = \mathbf{4 + 4} = \underline{\underline{\mathbf{8}}}$

Example 9.04

Pistons used in a certain type of engine have diameters that are known to follow a normal distribution with population mean 22.40 cm and population standard deviation 0.03 cm. Cylinders for this type of engine have diameters that are also distributed normally, with mean 22.50 cm and standard deviation 0.04 cm.

What is the probability that a randomly chosen piston will fit inside a randomly chosen cylinder?

Let X_P = piston diameter $\sim N(22.40, (0.03)^2)$
and X_C = cylinder diameter $\sim N(22.50, (0.04)^2)$

The random selection of piston and cylinder $\Rightarrow$ X_P and X_C are independent.

The chosen piston fits inside the chosen cylinder iff $X_P < X_C$.

$\Rightarrow \quad X_P - X_C < 0$

$E[X_P - X_C] = 22.40 - 22.50 = -0.10$

Independence $\Rightarrow$ $V[X_P - X_C] = V[X_P] + V[X_C]$
$= 0.000\,9 + 0.001\,6 = 0.002\,5$

$\Rightarrow \quad X_P - X_C \sim N(-0.10, (0.05)^2)$

$$P[X_P - X_C < 0] = P\left[Z < \frac{0-(-0.10)}{0.05}\right]$$

$$= \Phi(2.00)$$

$$= \underline{.977\,2} \quad \text{(to 4 s.f.)}$$

[It is very likely that a randomly chosen piston will fit inside a randomly chosen cylinder.]

[Follow-up exercise: with all other parameters unchanged, how small must the mean piston diameter be, so that P[fit] increases to 99%?

Answer:
$\mu_P = \mu_C - z_{.01} \times \sqrt{(V[X_P - X_C])} \approx 22.50 - 2.326 \times 0.05 = 22.38$ **cm (2 d.p.)]**

Distribution of the Sample Mean

If a random sample of size n is taken and the observed values are $\{X_1, X_2, X_3, \ldots, X_n\}$.

then the X_i are **independent and identically distributed** (***iid***) (each with population mean μ and population variance σ^2) and two more random quantities can be defined:

Sample total:

$$T = \sum_i X_i$$

Sample mean:

$$\overline{X} = \frac{T}{n} = \frac{1}{n}\sum_i X_i$$

$$\mathrm{E}\left[\overline{X}\right] = \frac{1}{n}\sum_i \mathrm{E}\left[X_i\right] = \frac{\mu}{n}\sum_i 1 = \frac{\mu}{n}\cdot n = \mu = \mathrm{E}\left[X\right]$$

Therefore $\boxed{\mathrm{E}\left[\overline{X}\right] = \mu}$

Also

$$\mathrm{V}\left[\overline{X}\right] = \mathrm{V}\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n^2}\mathrm{V}\left[\sum_i X_i\right]$$

$$= \frac{1}{n^2}\sum_i \mathrm{V}\left[X_i\right] \quad \left(\because \ \{X_i\} \text{ are independent}\right)$$

$$= \frac{1}{n^2}\sum_i \sigma^2 = \frac{\sigma^2}{n^2}\sum_i 1 = \frac{\sigma^2}{n^2}\cdot n$$

Therefore $\boxed{\mathrm{V}\left[\overline{X}\right] = \frac{\sigma^2}{n}}$

$\Rightarrow$ as $n \to \infty$, $\overline{X} \to \mu$.

Sampling Distributions

If $\quad X_i \sim N(\mu, \sigma^2)$ then $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

and $\quad Z = \dfrac{\overline{X} - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} \sim N(0, 1)$. $\qquad \left(\dfrac{\sigma}{\sqrt{n}}\right)$ is the **standard error**.

If σ^2 is unknown, then estimate it using $E[S^2] = \sigma^2$. Upon substitution into the expression for the standard normal random quantity Z , the additional uncertainty in σ changes the probability distribution of the random quantity from Z to

$$T = \frac{\overline{X} - \mu}{\left(\dfrac{S}{\sqrt{n}}\right)} \sim t_{n-1}, \quad \text{(a } t\text{-distribution with } \nu = (n-1) \text{ degrees of freedom)}$$

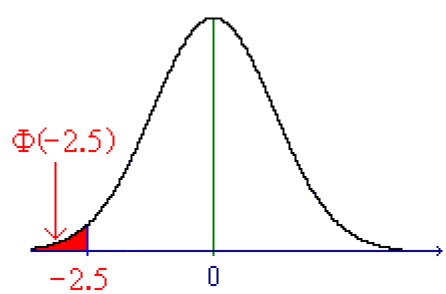
But $t_{n-1} \to N(0, 1)$ as $n \to \infty$.
[We shall employ the t distribution in Chapter 10].

Example 9.05

How likely is it that the mean of a random sample of 25 items, drawn from a normal population (with population mean and variance both equal to 100), will be less than 95?

$$P\left[\overline{X} < 95\right] = P\left[Z < \frac{\overline{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right] = P\left[Z < \frac{95 - 100}{\left(\frac{\sqrt{100}}{\sqrt{25}}\right)}\right]$$

$$= \Phi\left(\frac{-5}{\left(\frac{10}{5}\right)}\right) = \Phi(-2.50)$$



$\therefore \quad P\left[\overline{X} < 95\right] \approx \underline{\underline{.0062}}$

If $\overline{X} < 95$ occurs, then we may infer that "$X \sim N(100, 10^2)$" is *false*.

[However, in saying this, there remains a 0.62% chance that this inference is incorrect.]

Example 9.06

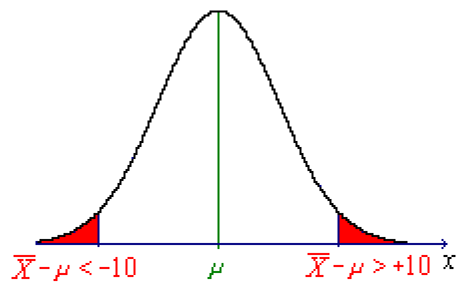
The mass of a sack of flour is normally distributed with a population mean of 1000 g and a population standard deviation of 35 g. Find the probability that the mean of a random sample of 25 sacks of flour will differ from the population mean by more than 10 g.

We require $\quad \mathrm{P}\left[\left|\overline{X}-\mu\right|>10\right]$.

$X \sim \mathrm{N}(1000, (35)^2) \qquad$ and $\qquad$ sample size $n = 25$

$$\Rightarrow \quad \overline{X} \sim \mathrm{N}\left(1000, \frac{35^2}{25}\right)$$

$$\Rightarrow \quad \text{standard error} = \frac{35}{5} = 7$$



$$\mathrm{P}\left[\left|\overline{X}-\mu\right| > 10\right]$$

$$= 2\times \mathrm{P}\left[\overline{X}-\mu < -10\right] \qquad (\because \text{ sym.})$$

$$= 2\times \mathrm{P}\left[Z < -\frac{10}{7}\right]$$

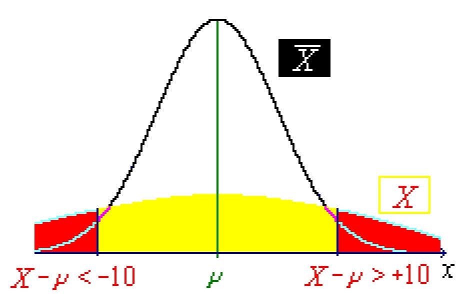
$$\approx \mathbf{2\,\Phi(-1.429)}$$

$$= \underline{\mathbf{.153}} \quad \textbf{(to 3 d.p.)}$$

Compare this with the probability of a single observation being at least that far away from the population mean:

$$\mathrm{P}\left[\left|X-\mu\right| > 10\right] = \ldots = 2\times\Phi\left(-\frac{10}{\boxed{35}}\right) = \underline{\underline{.772}} \quad (\text{to } 3 \text{ d.p.})$$

[The random quantity $\overline{X}$ is distributed much more tightly around μ than is any one individual observation X.]



Central Limit Theorem

If X_i is **not** normally distributed, but $\mathrm{E}[X_i] = \mu$, $\mathrm{V}[X_i] = \sigma^2$ and n is large (approximately 30 or more), then, to a good approximation,

$$\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

At `"http://www.engr.mun.ca/~ggeorge/3423/demos/clt.exe"` is a QBasic demonstration program to illustrate how the sample mean approaches a normal distribution even for highly non-normal distributions of X. [A list of other demonstration programs is at `"http://www.engr.mun.ca/~ggeorge/3423/demos/"`.]

Consider the exponential distribution, whose p.d.f. (probability density function) is

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad (x \ge 0, \lambda > 0) \quad \Rightarrow \quad \mathrm{E}[X] = \frac{1}{\lambda}, \quad \mathrm{V}[X] = \frac{1}{\lambda^2}$$

It can be shown that the exact p.d.f. of the sample mean for sample size n is

$$f_{\bar{X}}(x;\lambda,n) = \frac{\lambda n (\lambda n x)^{n-1} e^{-\lambda n x}}{(n-1)!}, \quad (x \ge 0, \lambda > 0, n \in \mathbb{N})$$

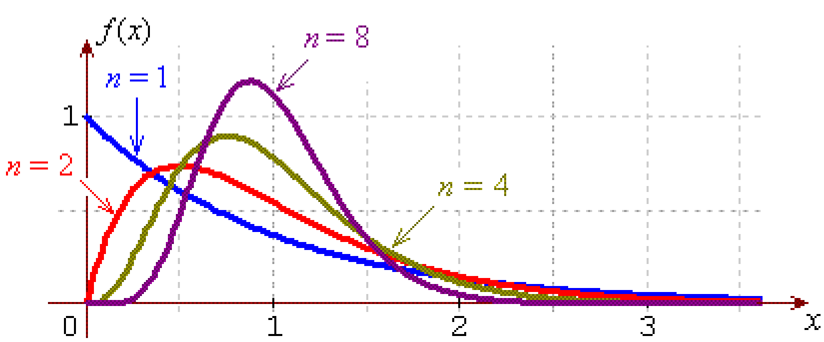
$$\Rightarrow \quad \mathrm{E}[\bar{X}] = \frac{1}{\lambda}, \quad \mathrm{V}[\bar{X}] = \frac{1}{n\lambda^2}$$

[A non-examinable derivation of this p.d.f. is available at `"http://www.engr.mun.ca/~ggeorge/3423/demos/cltexp2.doc"`.]

For illustration, setting $\lambda = 1$, the p.d.f. for the sample mean for sample sizes $n = 1, 2, 4$ and 8 are:

$n=1$: $\quad f(x) = e^{-x}$ $\qquad\qquad$ $n=2$: $\quad f_{\bar{X}}(x) = 4xe^{-2x}$

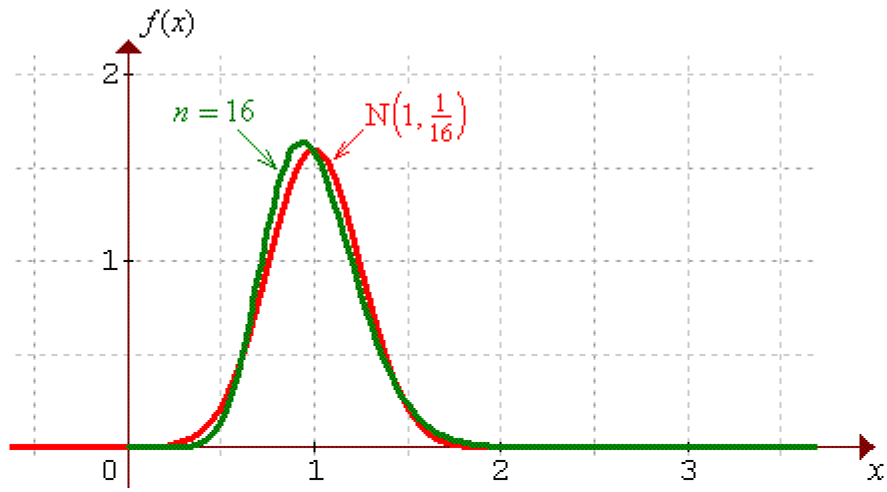
$n=4$: $\quad f_{\bar{X}}(x) = \dfrac{4(4x)^3 e^{-4x}}{3!}$ $\qquad\qquad$ $n=8$: $\quad f_{\bar{X}}(x) = \dfrac{8(8x)^7 e^{-8x}}{7!}$



The population mean $\mu = \mathrm{E}[X] = 1$ for all sample sizes.
The variance and the positive skew both diminish with increasing sample size.
The mode and the median approach the mean from the left.

For a sample size of $n = 16$, the sample mean $\overline{X}$ has the p.d.f.

$$f_{\overline{X}}(x) = \frac{16(16x)^{15} e^{-16x}}{15!}$$ and parameters $\mu = \mathrm{E}\left[\overline{X}\right] = 1$ and $\sigma^2 = \mathrm{V}\left[\overline{X}\right] = \frac{1}{16}$.



A plot of the exact p.d.f is drawn here, together with the normal distribution that has the same mean and variance. The approach to normality is clear. Beyond $n = 40$ or so, the difference between the exact p.d.f. and the Normal approximation is negligible.

It is generally the case that, whatever the probability distribution of a random quantity may be, the probability distribution of the sample mean $\overline{X}$ approaches normality as the sample size n increases. For most probability distributions of practical interest, the normal approximation becomes very good beyond a sample size of $n \approx 30$.

Example 9.07

A random sample of 100 items is drawn from an exponential distribution with parameter $\lambda = 0.04$. Find the probabilities that

(a) a single item has a value of more than 30;

(b) the sample mean has a value of more than 30.

(a)

$$\mathrm{P}[X > 30] = e^{-\lambda x} = e^{-.04 \times 30}$$
$$= e^{-1.2} = .301194\ldots$$
$$\approx \underline{\underline{\mathbf{.301}}}$$

(b)

$$\mu = \sigma = \frac{1}{\lambda} = \frac{1}{.04} = 25$$

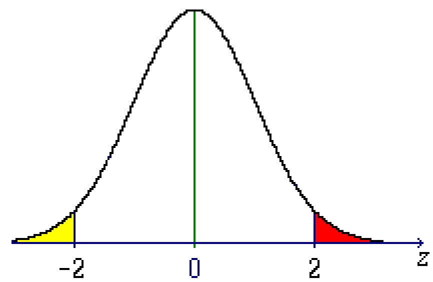
$$\Rightarrow \qquad \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{100}} = 2.5$$

***n* >> 30, so CLT $\Rightarrow$ $\overline{X} \sim$ N(25, (2.5)^2) to a good approximation.**

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} = \frac{30 - 25}{2.5} = 2$$

$$\therefore \qquad \mathrm{P}\left[\overline{X} > 30\right] = \mathrm{P}\left[Z > 2\right] = \Phi(-2.00)$$

$$\approx \underline{\mathbf{.0228}}$$



Sample Proportions

A Bernoulli random quantity has two possible outcomes:

$x = 0$ (= "failure") with probability $q = 1 - p$

and $x = 1$ (= "success") with probability p .

Suppose that all elements of the set $\{X_1, X_2, X_3, \ldots, X_n\}$ are independent Bernoulli random quantities, (so that the set forms a **random sample**).

Let $T = X_1 + X_2 + X_3 + \ldots + X_n$ = number of successes in the random sample

and $\hat{P} = \dfrac{T}{n}$ = proportion of successes in the random sample,

then T is **binomial (parameters: n, p)**

$\Rightarrow$ $\mathbf{E}[T] = np$

$\mathbf{V}[T] = npq$

$$\Rightarrow \quad \mathrm{E}\left[\hat{P}\right] = \mathrm{E}\left[\frac{1}{n}T\right] = \frac{np}{n} = p$$

$$\mathrm{V}\left[\hat{P}\right] = \mathrm{V}\left[\frac{1}{n}T\right] = \frac{npq}{n^2} = \frac{pq}{n}$$

For sufficiently large n, CLT $\Rightarrow$

$$\boxed{\hat{P} \sim \mathrm{N}\left(p, \frac{pq}{n}\right)}$$

Example 9.08

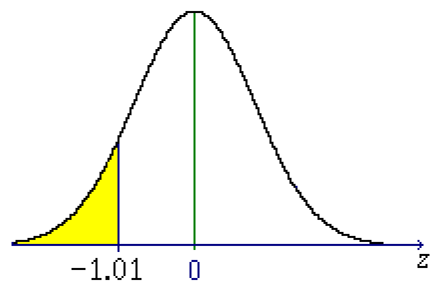
55% of all customers prefer brand *A*.
Find the probability that a majority in a random sample of 100 customers does **not** prefer brand *A*.

$$p = .55 \qquad n = 100$$

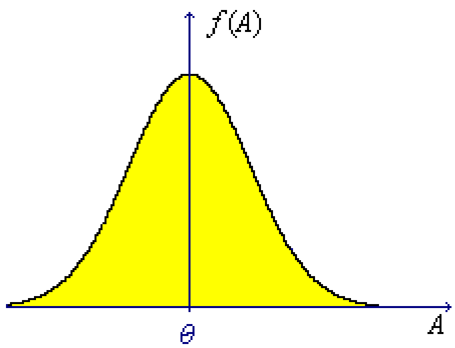
$$\Rightarrow \quad \hat{P} \sim \mathrm{N}\left(.55, \frac{.55 \times .45}{100}\right)$$

$$\Rightarrow \quad \mathrm{P}\left[\hat{P} < .50\right] = \mathrm{P}\left[Z < \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}\right] = \mathrm{P}\left[Z < \frac{.50 - .55}{\sqrt{.002\,475}}\right]$$

$$\approx \Phi(-1.01) \approx \underline{\underline{.156}}$$

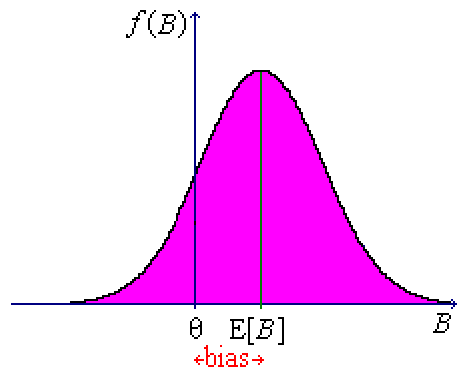


Unbiased estimator A for some unknown parameter θ:



$$\text{E}[A] = \theta$$

Biased estimator B for the unknown parameter θ:



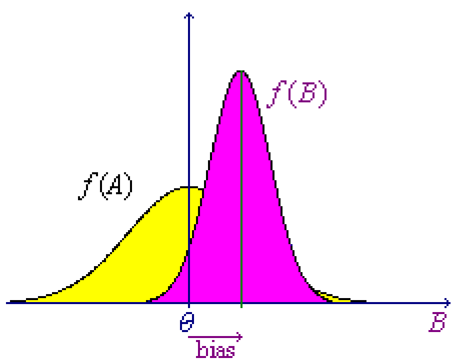
$$\text{E}[B] \neq \theta$$

Which estimator should we choose to estimate θ?

A is unbiased

BUT

B is more *efficient*
(accurate)



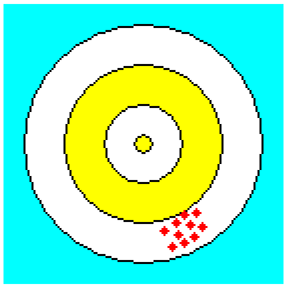
If P["B is closer than A to θ"] is high, then choose B,

else choose A.

A minimum variance unbiased estimator is ideal.
[See also Problem Set 6 Question 2]

Accuracy and Precision (Example 9.09)

An archer fires several arrows at the same target. **[The fable of William Tell, forced to use a cross-bow to shoot an apple off his son's head]**

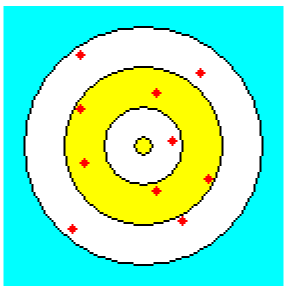


Precise

but

biased

[William is precise, but his son dies every time!]

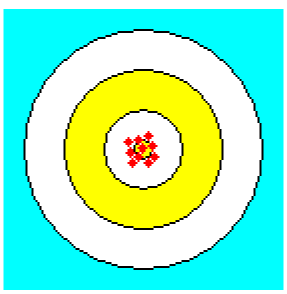


Unbiased

but

not precise

[William hits his son occasionally, often misses both son and apple, but, *on average*, is centred on the apple!]



Unbiased
and
precise

= accurate

[William hits the apple most of the time, to the relief of his son!]

Error = Systematic error + Random error

$(\text{inaccuracy})^2 = (\text{bias})^2 + \text{V[estimator]}$

Estimator A for θ is consistent iff

$$\text{E}[A] \to \theta \quad \text{and} \quad \text{V}[A] \to 0$$

(as $n \to \infty$)

A particular value a of an **estimator** A is an **estimate**.

Sample Mean

A random sample of n values $\{ X_1, X_2, X_3, \ldots, X_n \}$ is drawn from a population of mean μ and standard deviation σ.

Then $\mathrm{E}[X_i] = \mu$, $\mathrm{V}[X_i] = \sigma^2$ and the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$\overline{X}$ estimates μ .

$$\Rightarrow \quad \mathrm{E}\left[\overline{X}\right] = \mu , \qquad \mathrm{V}\left[\overline{X}\right] = \frac{\sigma^2}{n}$$

But, if μ is unknown, then σ^2 is unknown (usually).

Sample Variance

$$S^2 = \frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + \ldots + \left(X_n - \overline{X}\right)^2}{n-1}$$

$$= \frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n(n-1)}$$

and the **sample standard deviation** is $S = \sqrt{S^2}$

$n - 1$ = number of **degrees of freedom** for S^2 .

Justification for the divisor $(n - 1)$ [not examinable]:

Using $\mathrm{V}[Y] = \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2$ for all random quantities Y ,
$\mathrm{E}[Y^2] = \mathrm{V}[Y] + (\mathrm{E}[Y])^2 = \sigma_Y{}^2 + \mu_Y{}^2$

$$\Rightarrow \quad \mathrm{E}\left[\sum_i \left(X_i - \overline{X}\right)^2\right] = \mathrm{E}\left[\left(\sum_i X_i^2\right) - \frac{1}{n}\left(\sum_i X_i\right)^2\right]$$

$$= \mathrm{E}\left[\sum_i X_i^2\right] - \mathrm{E}\left[\frac{1}{n}\left(\sum_i X_i\right)^2\right] = \underbrace{\sum_i \mathrm{E}\left[X_i^2\right]}_{\text{set } Y = X_i} - \frac{1}{n}\underbrace{\mathrm{E}\left[\left(\sum_i X_i\right)^2\right]}_{\text{set } Y = \sum X_i}$$

$$= \left\{\sum_i \left(\mathrm{V}\left[X_i\right] + \left(\mathrm{E}\left[X_i\right]\right)^2\right)\right\} - \frac{1}{n}\left\{\mathrm{V}\left[\sum_i X_i\right] + \left(\mathrm{E}\left[\sum_i X_i\right]\right)^2\right\}$$

$$= \left\{\sum \left(\sigma^2 + \mu^2\right)\right\} - \frac{1}{n}\left\{n\sigma^2 + (n\mu)^2\right\} \qquad (\because \ i.i.d.)$$

$$= n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2} = (n-1)\sigma^2$$

Therefore

$$\mathrm{E}\left[\frac{\sum_i \left(X_i - \overline{X}\right)^2}{n-1}\right] = \sigma^2 \quad \Rightarrow \quad \mathrm{E}\left[S^2\right] = \sigma^2$$

but

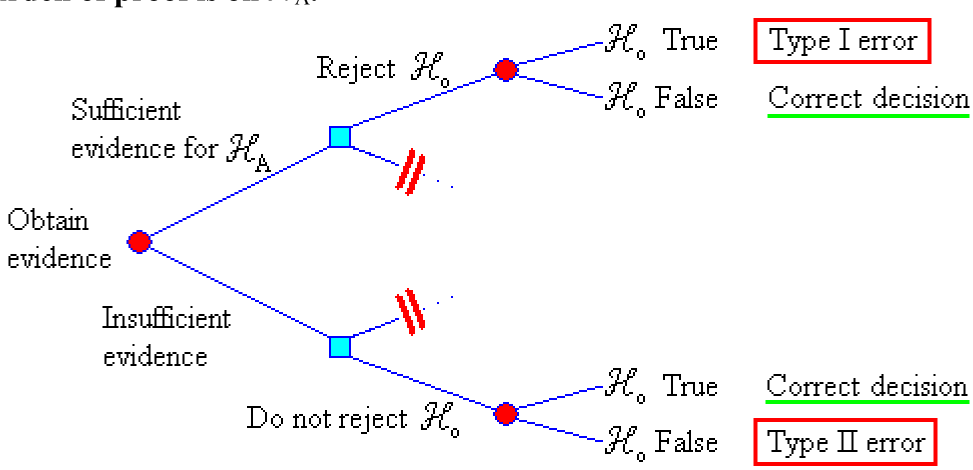
$$\mathrm{E}\left[\frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2\right] = \left(\frac{n-1}{n}\right)\sigma^2 < \sigma^2 \qquad \text{- biased!}$$

S^2 is the minimum variance unbiased estimator of σ^2 and
$\overline{X}$ is the minimum variance unbiased estimator of μ.
Both estimators are also consistent.

Inference – Some Initial Considerations

Is a **null hypothesis** $\mathcal{H}_o$ true (our "default belief"), or do we have sufficient evidence to reject $\mathcal{H}_o$ in favour of the **alternative hypothesis** $\mathcal{H}_A$?

$\mathcal{H}_o$ could be "defendant is not guilty" or "$\mu = \mu_o$" , etc.
The corresponding $\mathcal{H}_A$ could be "defendant is guilty" or "$\mu \neq \mu_o$" , etc.
The **burden of proof is on $\mathcal{H}_A$.**



Bayesian analysis: μ is treated as a random quantity. Data are used to modify prior belief about μ. Conclusions are drawn using both old and new information.
Classical analysis: Data are used to draw conclusions about μ, without using any prior information.