JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Review

Discrete RV's:

prob'y fctn: p(x) = Pr(X = x)
cdf: F(x) = Pr($X \leq$ x)
E(X) = $\sum_x$ x p(x)
SD(X) = $\sqrt{\text{E} \{ (X - \text{E } X)^2 \}}$

Continuous RV's:

density fctn: f(x)
cdf: F(x) = Pr($X \leq$ x)
E(X) = $\int$ x p(x) dx
SD(X) = $\sqrt{\text{E} \{ (X - \text{E } X)^2 \}}$

If Y = a + b X, then

E(Y) = a + b E(X) and SD(Y) = $|$b$|$ SD(X).

Example: if Z = (X – EX) / SD(X), then E(Z) = 0 and SD(Z) = 1

Review

Binomial(n,p):

no. successes in n indep. trials where
Pr(success) = p in each trial
If $X \sim$ binomial(n,p), then:
Pr(X = x) = $\binom{n}{p} p^x (1-p)^{n-x}$
E(X) = n p; SD(X) = $\sqrt{np(1-p)}$
E(X/n) = p; SD(X/p) = $\sqrt{p(1-p)/n}$

Poisson(λ):

Like a binomial(n,p), when n is very large
and p is very small. (λ = n p).
If $X \sim$ Poisson(λ), then:
Pr(X = x) = $e^{-\lambda} \lambda^x / x!$
E(X) = λ; SD(X) = $\sqrt{\lambda}$.

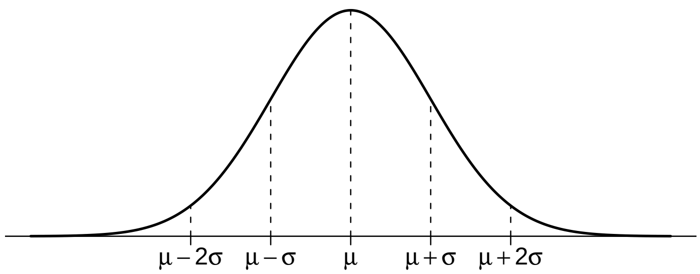
Normal distribution

If $X \sim N(\mu,\sigma)$,

density: $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$E(X) = \mu$ and $SD(X) = \sigma$

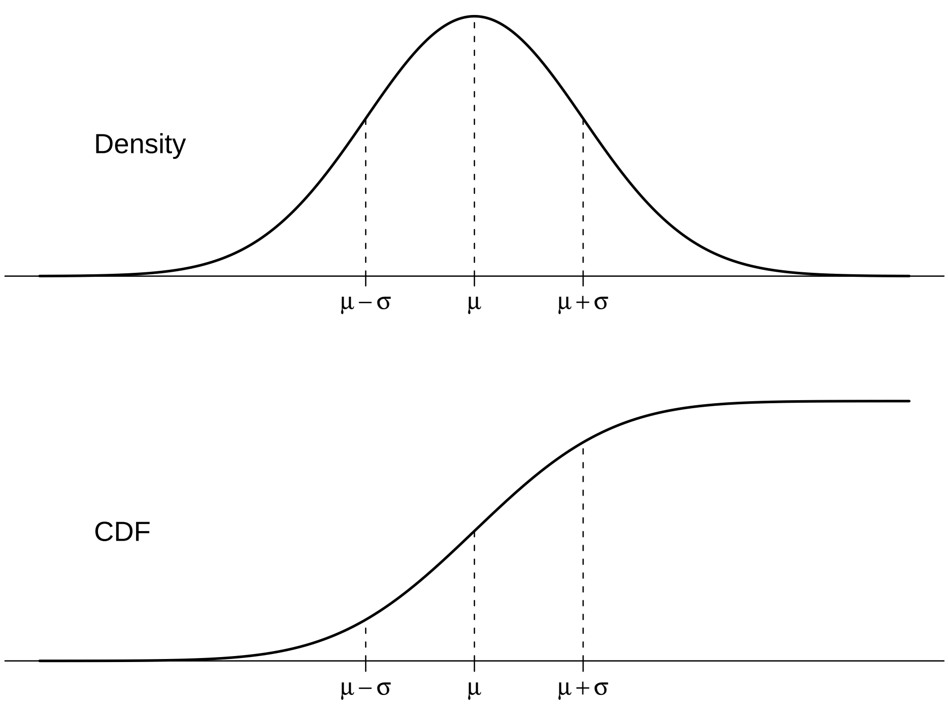
If $Z = (X - \mu) / \sigma$, then $Z \sim N(0,1)$ (the standard normal distr'n)



$Pr(\mu - \sigma \le X \le \mu + \sigma) \approx 68\%$

$Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 95\%$

The normal CDF



Calculations with the normal curve

In R:

- Convert to a statement involving the cdf
- Use the function `pnorm`

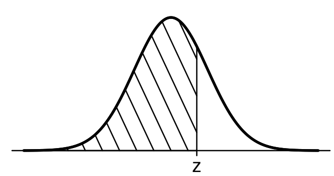
With a table:

- Convert to a statement involving the standard normal
- Convert to a statement involving the tabulated areas
- Look up the values in the table

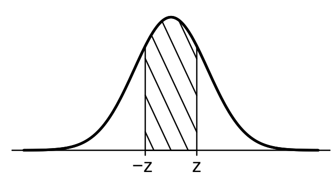
Draw a picture!

The tabulated areas

R



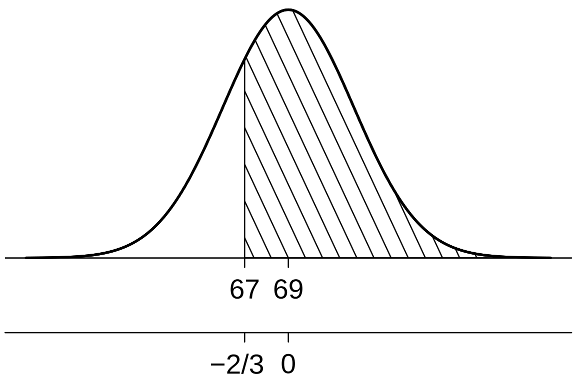
FPP table:



Examples

Suppose the heights of adult males in the U.S. are approximately normal distributed, with mean = 69 in and SD = 3 in.

What proportion of men are taller than 5'7"?

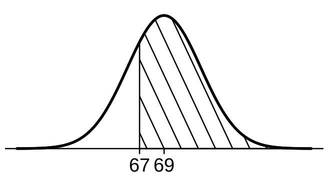


$X \sim N(\mu{=}69,\ \sigma{=}3)$

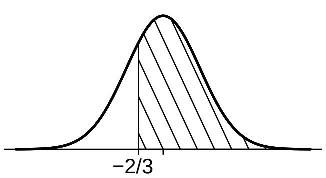
$Z = (X - 69)/3 \sim N(0,1)$

$$\begin{aligned}\Pr(X \geq 67) &= \Pr(Z \geq (67 - 69)/3) \\ &= \Pr(Z \geq -2/3)\end{aligned}$$

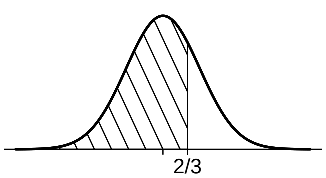
R



=



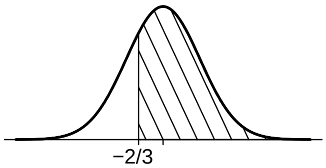
=



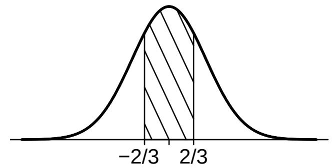
Use either `pnorm(2/3)` or `1 - pnorm(67, 69, 3)` or
`pnorm(67, 69, 3, lower=FALSE)`

The answer: 75%.

FPP table



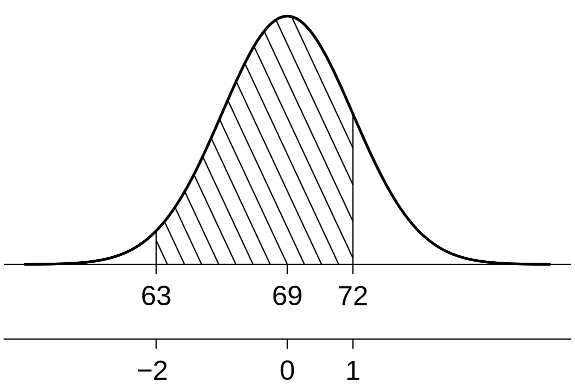
$= \frac{1}{2} + \frac{1}{2} \times$



$\approx 50\% + 48.43\% / 2 \approx 74\%$

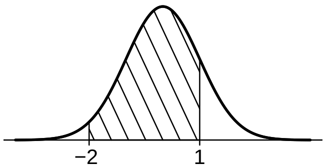
Another calculation

What proportion of men are between 5’3” and 6’?

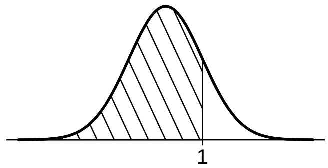


$\Pr(63 \leq X \leq 72) = \Pr(-2 \leq Z \leq 1)$

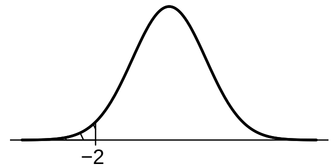
R



=



−




```
pnorm(72,69,3) - pnorm(63,69,3)
```

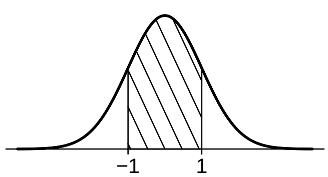
or

```
pnorm(1) - pnorm(-2)
```

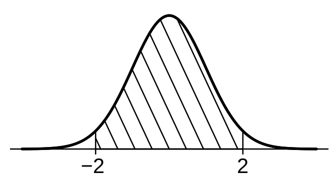
The answer: 82%.

FPP table

$\ldots \quad = \frac{1}{2} \times$



$+ \quad \frac{1}{2} \times$



$$\approx \frac{1}{2} \{ 68.27\% + 95.45\% \} \approx 82\%$$

Multiple Random Variables

We essentially always consider multiple RV's at once.

Key concepts: Joint, conditional and marginal distributions, and independence of RV's.

Let X and Y be discrete random variables.

Joint distribution:

$$p_{XY}(x,y) = \Pr(X = x \text{ and } Y = y)$$

Marginal distributions:

$$p_X(x) = \Pr(X = x) = \sum_y p_{XY}(x,y)$$
$$p_Y(y) = \Pr(Y = y) = \sum_x p_{XY}(x,y)$$

Conditional distributions:

$$p_{X|Y=y}(x) = \Pr(X = x \mid Y = y) = p_{XY}(x,y) / p_Y(y)$$

Example

Sample a couple who are both carriers of some disease gene

X = no. children they have

Y = no. affected children they have

		x						
	$p_{XY}(x,y)$	0	1	2	3	4	5	$p_Y(y)$
	0	0.160	0.248	0.124	0.063	0.025	0.014	0.634
	1	0	0.082	0.082	0.063	0.034	0.024	0.285
y	2	0	0	0.014	0.021	0.017	0.016	0.068
	3	0	0	0	0.003	0.004	0.005	0.012
	4	0	0	0	0	0.000	0.001	0.001
	5	0	0	0	0	0	0.000	0.000
	$p_X(x)$	0.160	0.330	0.220	0.150	0.080	0.060	

$\Pr(Y = y \mid X = 2)$

$p_{XY}(x,y)$		0	1	2	3	4	5	$p_Y(y)$
y	0	0.160	0.248	0.124	0.063	0.025	0.014	0.634
	1	0	0.082	0.082	0.063	0.034	0.024	0.285
	2	0	0	0.014	0.021	0.017	0.016	0.068
	3	0	0	0	0.003	0.004	0.005	0.012
	4	0	0	0	0	0.000	0.001	0.001
	5	0	0	0	0	0	0.000	0.000
$p_X(x)$		0.160	0.330	0.220	0.150	0.080	0.060	

y	0	1	2	3	4	5
$\Pr(Y=y \mid X=2)$	0.564	0.373	0.064	0.000	0.000	0.000

$\Pr(X = x \mid Y = 1)$

$p_{XY}(x,y)$		0	1	2	3	4	5	$p_Y(y)$
y	0	0.160	0.248	0.124	0.063	0.025	0.014	0.634
	1	0	0.082	0.082	0.063	0.034	0.024	0.285
	2	0	0	0.014	0.021	0.017	0.016	0.068
	3	0	0	0	0.003	0.004	0.005	0.012
	4	0	0	0	0	0.000	0.001	0.001
	5	0	0	0	0	0	0.000	0.000
$p_X(x)$		0.160	0.330	0.220	0.150	0.080	0.060	

x	0	1	2	3	4	5
$\Pr(X=x \mid Y=1)$	0.000	0.288	0.288	0.221	0.119	0.084

Independence

Random variables *X* and *Y* are independent if:

$p_{XY}(x,y) = p_X(x)\, p_Y(y)$ for every pair x,y

In other words/symbols:

$\Pr(X = x \text{ and } Y = y) = \Pr(X = x)\Pr(Y = y)$ for every pair x,y

Equivalently,

$\Pr(X = x \mid Y = y) = \Pr(X = x)$ for all x,y

Example

Sample a random rat from Baltimore.

X = 1 if the rat is infected with virus A, and = 0 otherwise

Y = 1 if the rat is infected with virus B, and = 0 otherwise

		x		
$p_{XY}(x,y)$		0	1	$p_Y(y)$
y	0	0.72	0.18	0.90
	1	0.08	0.02	0.10
	$p_X(x)$	0.80	0.20	

Continuous random variables

Continuous random variables have joint densities, say $f_{XY}(x,y)$.

The marginal densities are obtained by integration:

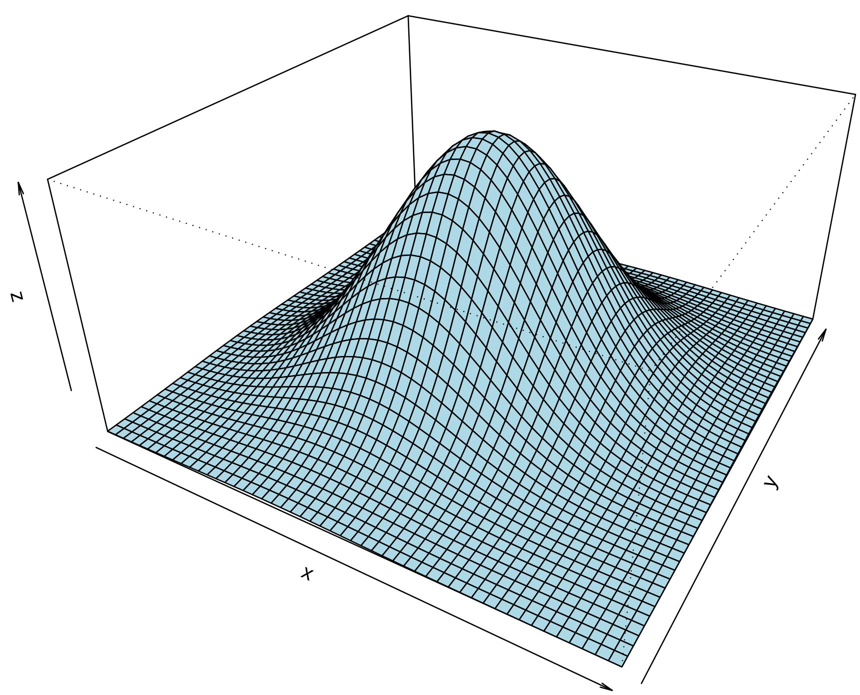
$$f_X(x) = \int f_{XY}(x,y)\,dy \quad \text{and} \quad f_Y(y) = \int f_{XY}(x,y)\,dx$$

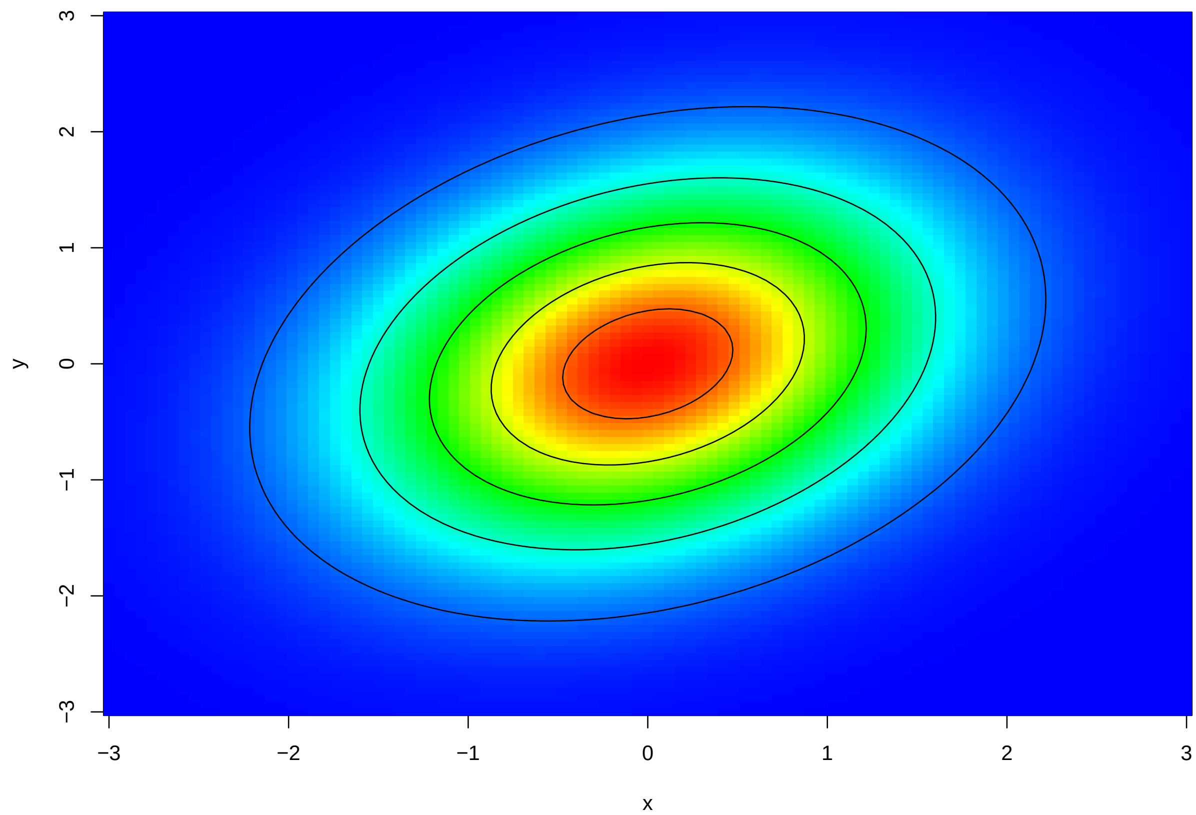
Conditional density:

$$f_{X|Y=y}(x) = f_{XY}(x,y)/f_Y(y)$$

X and Y are independent if

$$f_{XY}(x,y) = f_X(x)\,f_Y(y) \qquad \text{for all } x,y$$





iid

More jargon:

Random variables $X_1, X_2, X_3, \ldots, X_n$ are said to be independent and identically distributed (iid) if:

(a) they are independent and

(b) they all have the same distribution

Usually:

- Repeated independent measurements
- Random sampling from a large population

Means and SDs

Mean and SD of sums of random variables:

$E(\sum_i X_i) = \sum_i E(X_i)$ no matter what

$SD(\sum_i X_i) = \sqrt{\sum_i \{SD(X_i)\}^2}$ if the X_i are independent

Mean and SD of means of random variables:

$E(\sum_i X_i / n) = \sum_i E(X_i)/n$ no matter what

$SD(\sum_i X_i/n) = \sqrt{\sum_i \{SD(X_i)\}^2}/n$ if the X_i are independent

If the X_i are iid with mean μ and SD σ:

$E(\sum_i X_i / n) = \mu$ and $SD(\sum_i X_i / n) = \sigma/\sqrt{n}$