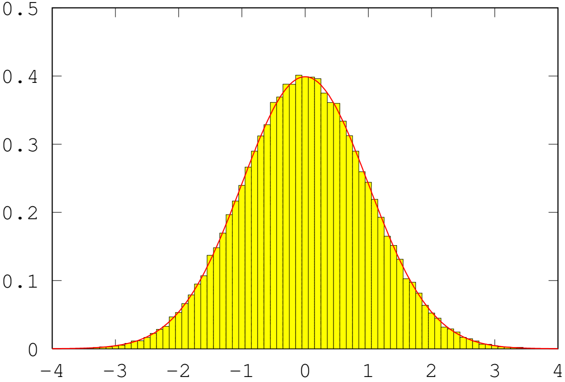
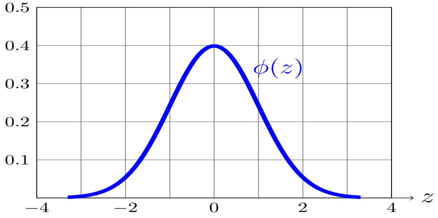
Central Limit Theorem, Joint Distributions
18.05 Spring 2018



Exam next Wednesday

- Exam 1 on Wednesday March 7, regular room and time.
- Designed for 1 hour. You will have the full 80 minutes.
- Class on Monday will be review.
- Practice materials posted.
- Learn to use the standard normal table for the exam.
- No books or calculators.
- You may have one 4×6 notecard with any information you like.

The bell-shaped curve



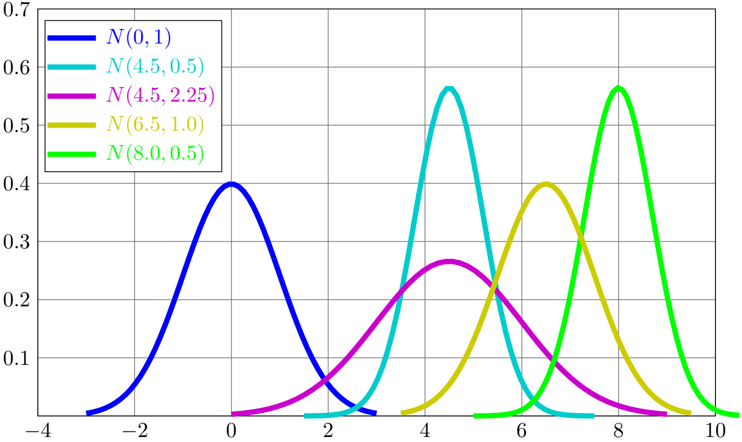
- This is standard normal distribution $N(0, 1)$:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

- $N(0, 1)$ means that mean is $\mu = 0$, and std deviation is $\sigma = 1$.
- Normal with mean μ, std deviation σ is $N(\mu, \sigma)$:

$$\phi_{\mu,\sigma}(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(z-\mu)^2/2\sigma^2}$$

Lots of normal distributions



Standardization

Random variable X with mean μ, standard deviation σ.

Standardization: $$Y = \frac{X - \mu}{\sigma}.$$

- Y has mean 0 and standard deviation 1.
- Standardizing any normal random variable produces the standard normal.
- If $X \approx$ normal then standardized $X \approx$ stand. normal.
- We reserve Z to mean a standard normal random variable.

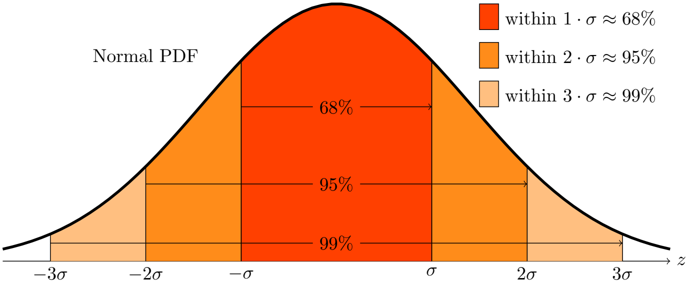
Board Question: Standardization

Here are the pdfs for four (binomial) random variables X. Standardize them, and make bar graphs of the standardized distributions. Each bar should have area equal to the probability of that value. (Each bar has width $1/\sigma$, so each bar has height pdf$\cdot\sigma$.)

X	$n = 0$	$n = 1$	$n = 4$	$n = 9$
0	1	1/2	1/16	1/512
1	0	1/2	4/16	9/512
2	0	0	6/16	36/512
3	0	0	4/16	84/512
4	0	0	1/16	126/512
5	0	0	0	126/512
6	0	0	0	84/512
7	0	0	0	36/512
8	0	0	0	9/512
9	0	0	0	1/512

Concept Question: Normal Distribution

X has normal distribution, standard deviation σ.



1. $P(-\sigma < X < \sigma)$ is

(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

2. $P(X > 2\sigma)$

(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

answer: 1c, 2a

Central Limit Theorem

Setting: $X_1, X_2, \ldots$ i.i.d. with mean μ and standard dev. σ.

For each n:

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \qquad \text{average}$$

$$S_n = X_1 + X_2 + \ldots + X_n \qquad \text{sum.}$$

Conclusion: For large n:

$$\overline{X}_n \approx \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

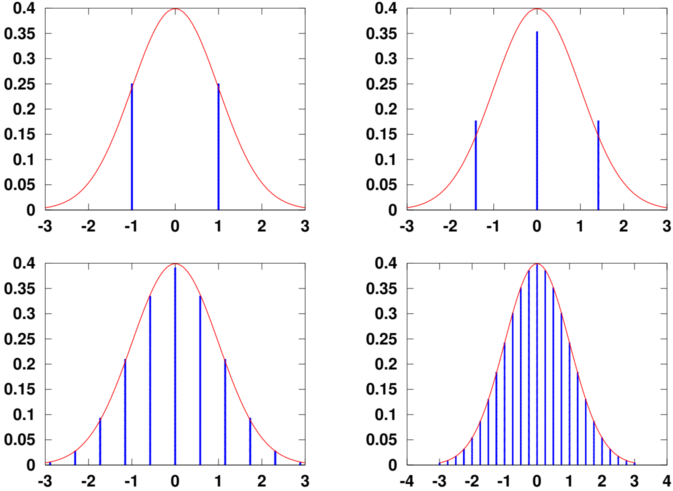
$$S_n \approx \mathrm{N}\left(n\mu, n\sigma^2\right)$$

$$\text{Standardized } \left(S_n \text{ or } \overline{X}_n\right) \approx \mathrm{N}(0,1)$$

$$\text{That is,} \quad \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx \mathrm{N}(0,1).$$

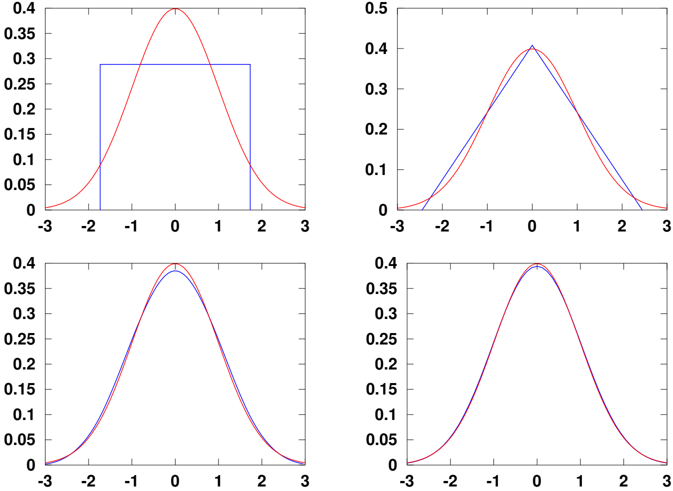
CLT: pictures

The standardized average of n i.i.d. Bernoulli(0.5) random variables with $n = 1, 2, 12, 64$.



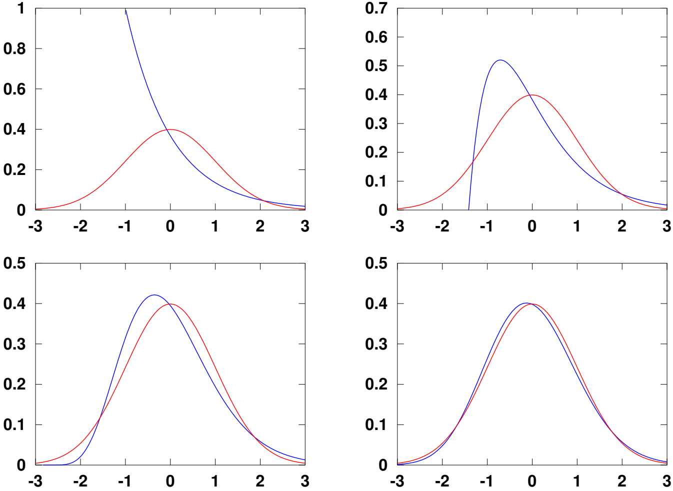
CLT: pictures 2

Standardized average of n i.i.d. uniform random variables with $n = 1, 2, 4, 12$.



CLT: pictures 3

The standardized average of n i.i.d. exponential random variables with $n = 1, 2, 8, 64$.



CLT: pictures

The non-standardized average of n Bernoulli(0.5) random variables, with $n = 4, 12, 64$. Spikier.

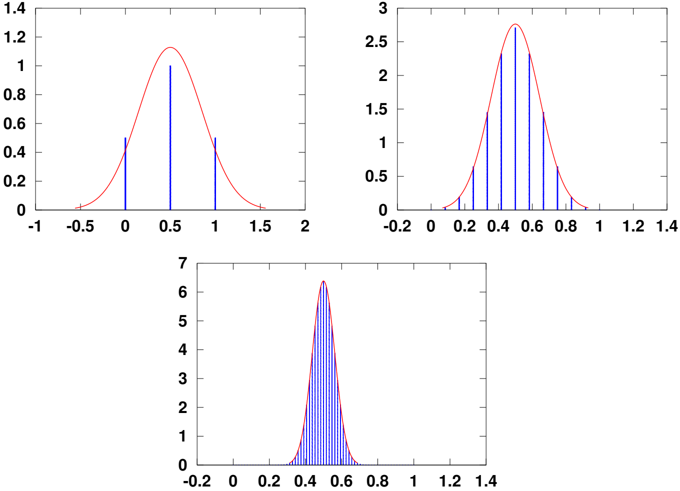


Table Question: Sampling from the standard normal distribution

As a table, produce two random samples from (an approximate) standard normal distribution.

To make each sample, the table is allowed eight rolls of the 10-sided die.

Note: $\mu = 5.5$ and $\sigma^2 \approx 8$ for a single 10-sided die.

Hint: CLT is about averages.

<u>**answer:**</u> The average of 9 rolls is a sample from the average of 9 independent random variables. The CLT says this average is approximately normal with $\mu = 5.5$ and $\sigma = 8.25/\sqrt{9} = 2.75$
If $\overline{x}$ is the average of 9 rolls then standardizing we get

$$z = \frac{\overline{x} - 5.5}{2.75}$$

is (approximately) a sample from $\mathrm{N}(0, 1)$.

Board Question: CLT

1. Carefully write the statement of the central limit theorem.

2. To head the newly formed US Dept. of Statistics, suppose that 50% of the population supports Ani, 25% supports Ruthi, and the remaining 25% is split evenly between Efrat, Elan, David and Jerry.

A poll asks 400 random people who they support. What is the probability that at least 55% of those polled prefer Ani?

3. What is the probability that less than 20% of those polled prefer Ruthi?

answer: On next slide.

Solution

answer: 2. Let $\mathcal{A}$ be the fraction polled who support Ani. So $\mathcal{A}$ is the average of 400 Bernoulli(0.5) random variables. That is, let $X_i = 1$ if the ith person polled prefers Ani and 0 if not, so $\mathcal{A}$ = average of the X_i.
The question asks for the probability $\mathcal{A} > 0.55$.

Each X_i has $\mu = 0.5$ and $\sigma^2 = 0.25$. So, $E(\mathcal{A}) = 0.5$ and $\sigma_{\mathcal{A}}^2 = 0.25/400$ or $\sigma_{\mathcal{A}} = 1/40 = 0.025$.

Because $\mathcal{A}$ is the average of 400 Bernoulli(0.5) variables the CLT says it is approximately normal and standardizing gives

$$\frac{\mathcal{A} - 0.5}{0.025} \approx Z$$

So

$$P(\mathcal{A} > 0.55) \approx P(Z > 2) \approx 0.025$$

Continued on next slide

Solution continued

3. Let $\mathcal{R}$ be the fraction polled who support Ruthi.
The question asks for the probability the $\mathcal{R} < 0.2$.
Similar to problem 2, $\mathcal{R}$ is the average of 400 Bernoulli(0.25) random variables. So

$$E(\mathcal{R}) = 0.25 \quad \text{and} \quad \sigma_{\mathcal{R}}^2 = (0.25)(0.75)/400 = \Rightarrow \ \sigma_{\mathcal{R}} = \sqrt{3}/80.$$

So $\dfrac{\mathcal{R} - 0.25}{\sqrt{3}/80} \approx Z$. So,

$$P(\mathcal{R} < 0.2) \approx P(Z < -4/\sqrt{3}) \approx 0.0105$$

Bonus problem

Not for class. Solution will be posted with the slides.
An accountant rounds to the nearest dollar. We'll assume the error in rounding is uniform on [-0.5, 0.5]. Estimate the probability that the total error in 300 entries is more than $5.

answer: Let X_j be the error in the j^{th} entry, so, $X_j \sim U(-0.5, 0.5)$.

We have $E(X_j) = 0$ and $\text{Var}(X_j) = 1/12$.

The total error $S = X_1 + \ldots + X_{300}$ has $E(S) = 0$,
$\text{Var}(S) = 300/12 = 25$, and $\sigma_S = 5$.

Standardizing we get, by the CLT, $S/5$ is approximately standard normal.
That is, $S/5 \approx Z$.

So $P(S < -5 \text{ or } S > 5) \approx P(Z < -1 \text{ or } Z > 1) \approx \boxed{0.32}$.

Joint Distributions

X and Y are jointly distributed random variables.

Discrete: Probability mass function (pmf):

$$p(x_i, y_j)$$

Continuous: probability density function (pdf):

$$f(x, y)$$

Both: cumulative distribution function (cdf):

$$F(x, y) = P(X \leq x, Y \leq y)$$

Discrete joint pmf: example 1

Roll two dice: $X = \#$ on first die, $Y = \#$ on second die

X takes values in 1, 2, ..., 6, Y takes values in 1, 2, ..., 6

Joint probability table:

$X\backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

pmf: $p(i, j) = 1/36$ for any i and j between 1 and 6.

Discrete joint pmf: example 2

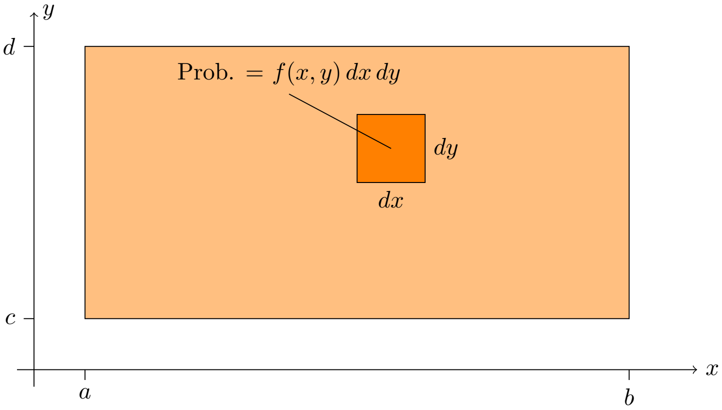
Roll two dice: $X = \#$ on first die, $T =$ total on both dice

$X\backslash T$	2	3	4	5	6	7	8	9	10	11	12
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36

Continuous joint distributions

- X takes values in $[a, b]$, $\qquad Y$ takes values in $[c, d]$
- (X, Y) takes values in $[a, b] \times [c, d]$.
- Joint probability density function (pdf) $\quad f(x, y)$

$f(x, y)\, dx\, dy$ is the probability of being in the small square.



Properties of the joint pmf and pdf

Discrete case: probability mass function (pmf)

1. $0 \le p(x_i, y_j) \le 1$
2. Total probability is 1:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

Continuous case: probability density function (pdf)

1. $0 \le f(x, y)$
2. Total probability is 1:

$$\int_c^d \int_a^b f(x, y)\, dx\, dy = 1$$

Note: $f(x, y)$ can be greater than 1: it is a density, *not* a probability.

Example: discrete events

Roll two dice: $X = \#$ on first die, $Y = \#$ on second die.

Consider the event: $A =$ '$Y - X \geq 2$'

Describe the event A and find its probability.

answer: We can describe A as a set of (X, Y) pairs:

$$A = \{(1,3),(1,4),(1,5),(1,6),(2,4),(2,5),(2,6),(3,5),(3,6),(4,6)\}.$$

Or we can visualize it by shading the table:

$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

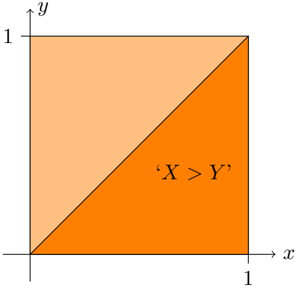
$P(A) =$ sum of probabilities in shaded cells $= 10/36$.

Example: continuous events

Suppose (X, Y) takes values in $[0,1] \times [0,1]$.

Uniform density $f(x, y) = 1$.

Visualize the event '$X > Y$' and find its probability.
answer:



The event takes up half the square. Since the density is uniform this is half the probability. That is, $P(X > Y) = 0.5$.

Cumulative distribution function

$$F(x,y) = P(X \le x,\ Y \le y) \ = \ \int_c^y \int_a^x f(u,v)\, du\, dv.$$

$$f(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y).$$

Properties

1. $F(x,y)$ is non-decreasing. That is, as x or y increases $F(x,y)$ increases or remains constant.
2. $F(x,y) = 0$ at the lower left of its range.
   If the lower left is $(-\infty, -\infty)$ then this means
   $$\lim_{(x,y)\to(-\infty,-\infty)} F(x,y) = 0.$$
3. $F(x,y) = 1$ at the upper right of its range.

Marginal pmf and pdf

Roll two dice: $X = \#$ on first die, $T =$ total on both dice.

The marginal pmf of X is found by summing the rows. The marginal pmf of T is found by summing the columns

$X \backslash T$	2	3	4	5	6	7	8	9	10	11	12	$p(x_i)$
1	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	0	1/6
2	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	0	1/6
3	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	0	1/6
4	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	0	1/6
5	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	0	1/6
6	0	0	0	0	0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(t_j)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	1

For continuous distributions the marginal pdf $f_X(x)$ is found by integrating out the y. Likewise for $f_Y(y)$.

Board question

Suppose X and Y are random variables and

- (X, Y) takes values in $[0,1] \times [0,1]$.
- the pdf is $\dfrac{3}{2}(x^2 + y^2)$.

1. Show $f(x, y)$ is a valid pdf.
2. Visualize the event $A =$ '$X > 0.3$ and $Y > 0.5$'. Find its probability.
3. Find the cdf $F(x, y)$.
4. Find the marginal pdf $f_X(x)$. Use this to find $P(X < 0.5)$.
5. Use the cdf $F(x, y)$ to find the marginal cdf $F_X(x)$ and $P(X < 0.5)$.
6. See next slide

Board question continued

6. (New scenario) From the following table compute $F(3.5, 4)$.

$X \backslash Y$	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

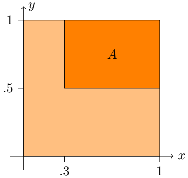
answer: See next slide

Solution

answer: 1. Validity: Clearly $f(x, y)$ is positive. Next we must show that total probability $= 1$:

$$\int_0^1 \int_0^1 \frac{3}{2}(x^2 + y^2)\, dx\, dy = \int_0^1 \left[\frac{1}{2}x^3 + \frac{3}{2}xy^2\right]_0^1 dy = \int_0^1 \frac{1}{2} + \frac{3}{2}y^2\, dy = 1.$$

2. Here's the visualization



The pdf is not constant so we must compute an integral

$$P(A) = \int_{.3}^1 \int_{.5}^1 \frac{3}{2}(x^2 + y^2)\, dy\, dx = \int_{.3}^1 \left[\frac{3}{2}x^2 y + \frac{1}{2}y^3\right]_{.5}^1 dx$$

(continued)

Solutions 2, 3, 4, 5

2. (continued) $\quad = \displaystyle\int_{.3}^{1} \frac{3x^2}{4} + \frac{7}{16}\, dx = \boxed{0.5495}$

3. $F(x, y) = \displaystyle\int_0^y \int_0^x \frac{3}{2}(u^2 + v^2)\, du\, dv = \boxed{\frac{x^3 y}{2} + \frac{xy^3}{2}}.$

4.

$$f_X(x) = \int_0^1 \frac{3}{2}(x^2 + y^2)\, dy = \left[\frac{3}{2}x^2 y + \frac{y^3}{2}\right]_0^1 = \boxed{\frac{3}{2}x^2 + \frac{1}{2}}$$

$$P(X < .5) = \int_0^{.5} f_X(x)\, dx = \int_0^{.5} \frac{3}{2}x^2 + \frac{1}{2}\, dx = \left[\frac{1}{2}x^3 + \frac{1}{2}x\right]_0^{.5} = \boxed{\frac{5}{16}}.$$

5. To find the marginal cdf $F_X(x)$ we simply take y to be the top of the y-range and evalute F: $\quad F_X(x) = F(x, 1) = \dfrac{1}{2}(x^3 + x).$

Therefore $\quad P(X < .5) = F(.5) = \dfrac{1}{2}(\dfrac{1}{8} + \dfrac{1}{2}) = \boxed{\dfrac{5}{16}}.$

6. On next slide

Solution 6

6. $F(3.5, 4) = P(X \leq 3.5, Y \leq 4)$.

$X \backslash Y$	1	2	3	4	5	6
1	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
2	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
3	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
4	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
5	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
6	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$

Add the probability in the shaded squares: $F(3.5, 4) = 12/36 = 1/3$.