4.3 Normal distribution

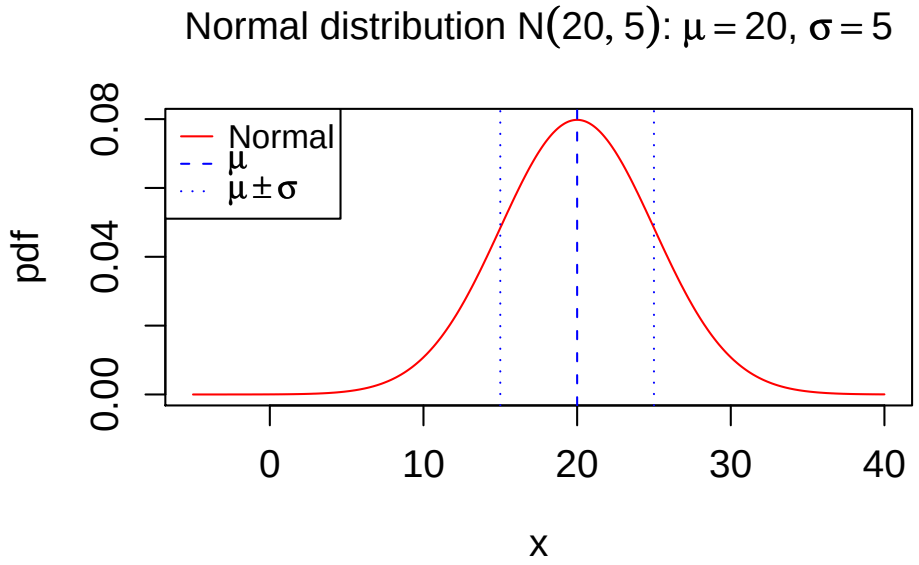
Prof. Tesler

Math 186
Winter 2020

Normal distribution

a.k.a. “Bell curve” and “Gaussian distribution”

- The *normal distribution* is a continuous distribution. Parameters:
  μ = mean (center)
  σ = standard deviation (width)
- PDF: $$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{for } -\infty < x < \infty.$$



- The normal distribution is symmetric about $x = \mu$, so median = mean = μ.

Applications of normal distribution

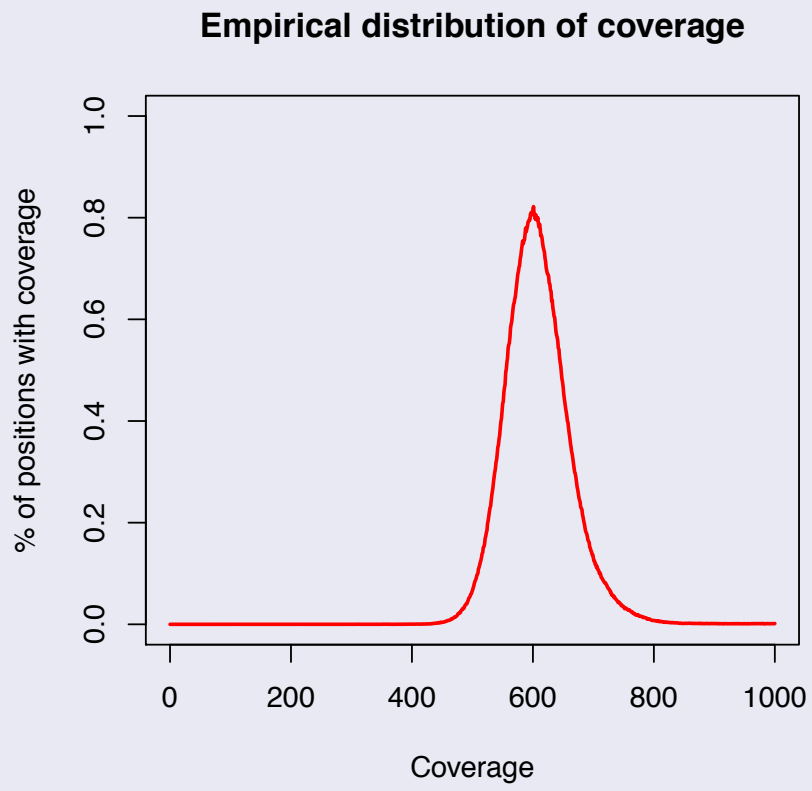
Applications

- Many natural quantities are modelled by it: e.g., a histogram of the heights or weights of everyone in a large population often follows a normal distribution.
- Many distributions such as binomial, Poisson,... are closely approximated by it when the parameters are large enough.
- Sums and averages of huge quantities of data are often modelled by it.

Coverage in DNA sequencing

Illumina GA_{II} sequencing of *E. coli* at 600$\times$ coverage.

Chitsaz et al. (2011), *Nature Biotechnology*



Cumulative distribution function

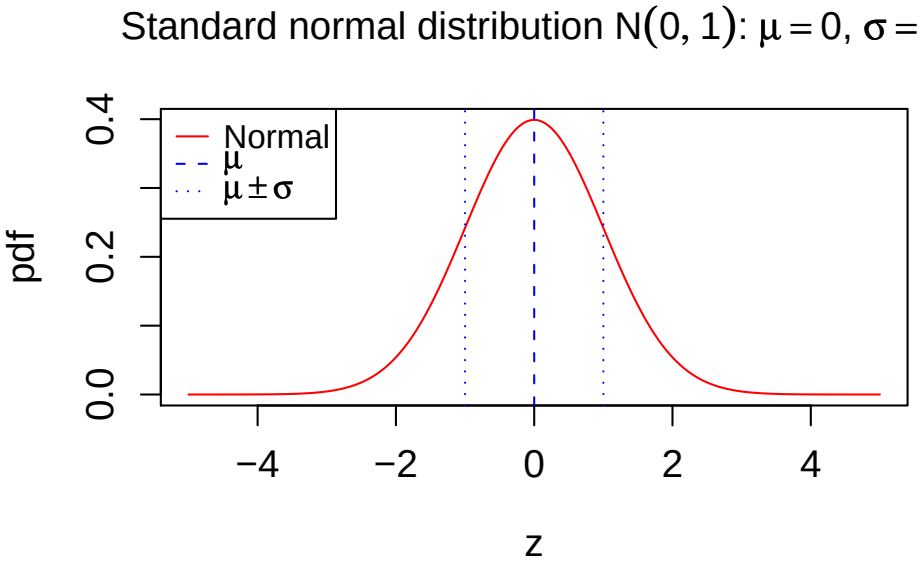
- The cumulative distribution function is the integral

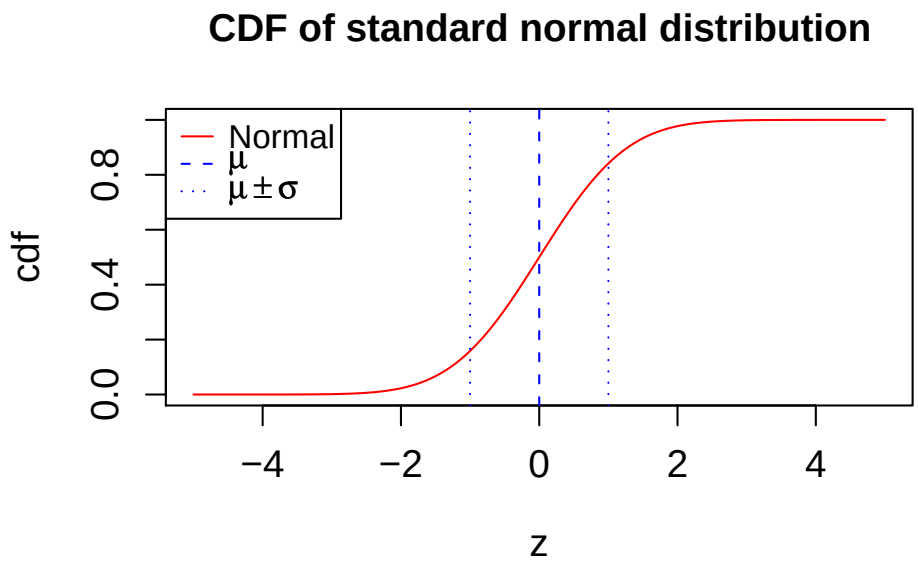
$$F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\, dt$$

- The usual strategy to compute integrals is antiderivatives, like $\int x^2\, dx = \frac{x^3}{3} + C$. But this doesn't have an antiderivative in terms of the usual functions (polynomials, exponentials, logs, trig, ...).
- It can be done via numerical integration or Taylor series.
- We'll learn how to do it with a look-up table.
- The integral for total probability equals 1; this can be shown using double integrals in polar coordinates:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\, dx = 1$$

Standard normal distribution





- The *standard normal distribution* is the normal distribution for $\mu = 0$, $\sigma = 1$. Use the variable name Z:

$$\text{PDF:} \quad \phi(z) = f_Z(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \qquad \text{for } -\infty < z < \infty$$

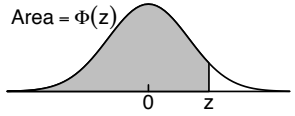
$$\text{CDF:} \quad \Phi(z) = F_Z(z) = P(Z \leqslant z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

- Compute $\Phi(z)$ with the table in the book (Table A.1 in the back).

Standard normal distribution — tables

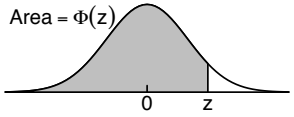
Table A.1 in the back of the book is similar to this

Cumulative Area Under the Standard Normal Distribution



z	0	1	2	3	4	5	6	7	8	9
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Cumulative Area Under the Standard Normal Distribution

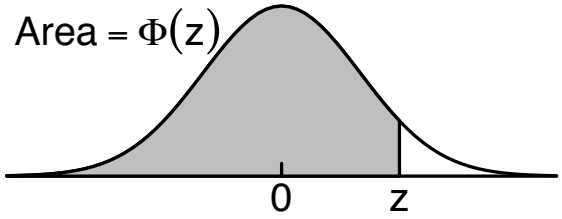


z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Standard normal distribution — tables

Table A.1 in the back of the book is similar to this

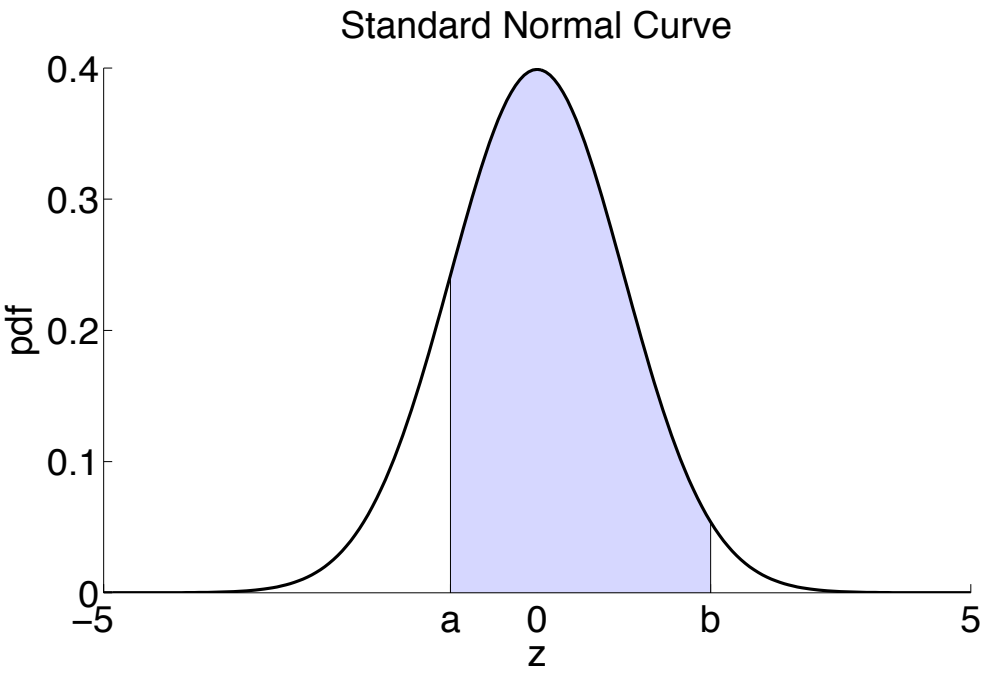
Cumulative Area Under the Standard Normal Distribution



z	0	1	2	3	4	5	6	7	8	9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	**0.0655**	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	**0.9345**	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	**0.9474**	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$\Phi(1.51) \approx 0.9345 \qquad \Phi(1.62) \approx 0.9474 \qquad \Phi(-1.51) \approx 0.0655$$

Standard normal distribution — areas

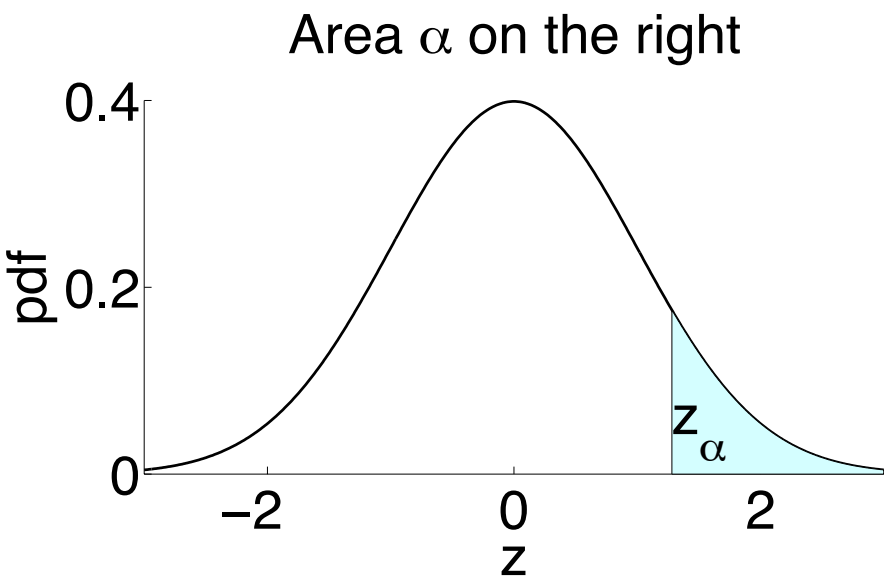


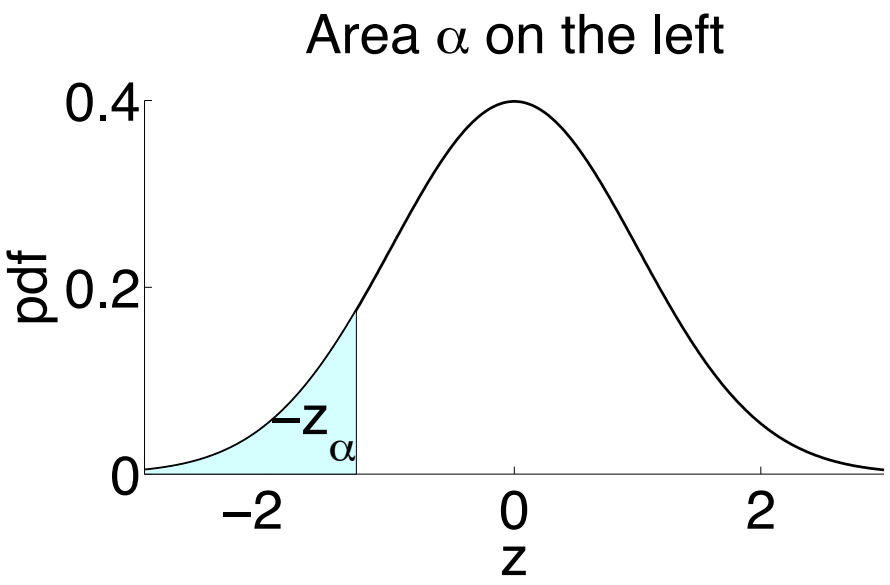
- The area between $z = a$ and $z = b$ is

$$P(a \leqslant Z \leqslant b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt = \Phi(b) - \Phi(a)$$

- Use table in back of book:
  $P(1.51 \leqslant Z \leqslant 1.62) = \Phi(1.62) - \Phi(1.51) = 0.9474 - 0.9345 = 0.0129$

Standard normal distribution — symmetries of areas





- Area right of z is $P(Z > z) = 1 - \Phi(z)$.
- By symmetry, the area left of $-z$ and the area right of z are equal:

$$\Phi(-z) = 1 - \Phi(z)$$

$$\Phi(-1.51) = 1 - \Phi(1.51) = 1 - 0.9345 = 0.0655$$

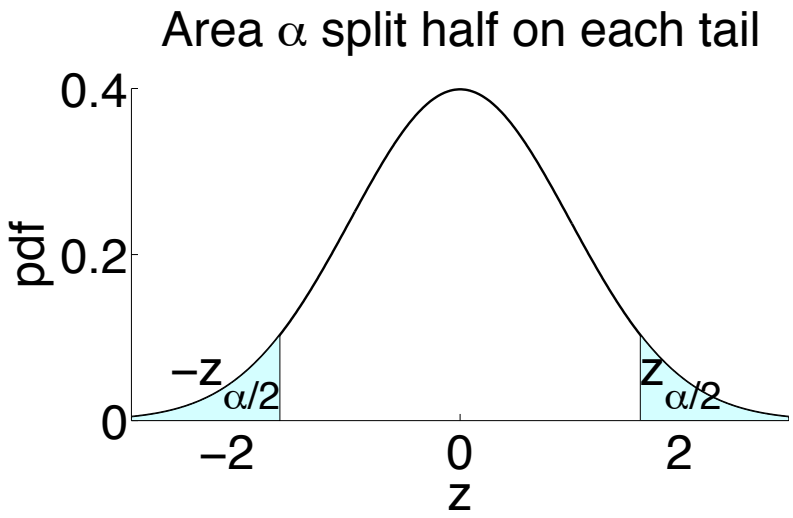
- Area between $z = \pm a$:

$$\Phi(a) - \Phi(-a) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1$$

$$\Phi(1.51) - \Phi(-1.51) = 2\Phi(1.51) - 1 \approx .8690$$

Central area

- Area between $z = \pm 1$:
  $\Phi(1) - \Phi(-1) = 0.6827 = 68.27\%$
  Area between $z = \pm 1$ is $\approx 68.27\%$.
- Area between $z = \pm 2$ is $\approx 95.45\%$.
- Area between $z = \pm 3$ is $\approx 99.73\%$.



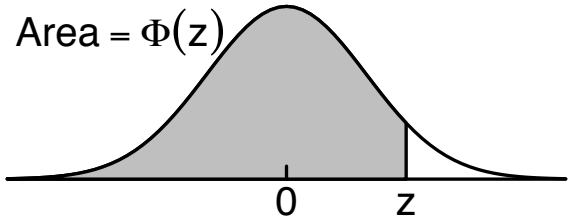
Find the center part containing 95% of the area

- Put 2.5% of the area at the upper tail, 2.5% at the lower tail, and 95% in the middle.
- The value of z putting 2.5% at the top gives
  $\Phi(z) = 1 - 0.025 = 0.975$.
- Using the table in the book, $z \approx 1.96$.
- **Notation:** $z_{.025} = 1.96$. The area between $z = \pm 1.96$ is about 95%.
- For 99% in the middle, 0.5% on each side, use $z_{.005} \approx 2.58$.

Central area

Find z with $\Phi(z) \approx 0.9750$

Cumulative Area Under the Standard Normal Distribution



z	0	1	2	3	4	5	6	7	8	9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	**0.0250**	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	**0.9750**	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$\Phi(1.96) \approx 0.9750 \qquad \Phi(-1.96) \approx 0.0250$$

$$z_{0.025} = 1.960$$

Areas on normal curve for arbitrary μ, σ

$$P(a \leqslant X \leqslant b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\, dx$$

- Substitute $z = \frac{x-\mu}{\sigma}$ (or $x = \sigma z + \mu$) into the x integral to turn it into the standard normal integral:

$$P\left(\frac{a-\mu}{\sigma} \leqslant \frac{X-\mu}{\sigma} \leqslant \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \leqslant Z \leqslant \frac{b-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

- The *z-score* of x is $z = \frac{x-\mu}{\sigma}$.

Binomial distribution

Compute $P(43 \leqslant X \leqslant 51)$ when $n = 60$, $p = 3/4$

Binomial: $n = 60$, $p = 3/4$

k	$P(X = k) = \binom{60}{k}(.75)^k(.25)^{60-k}$
43	0.09562
44	0.11083
45	0.11822
46	0.11565
47	0.10335
48	0.08397
49	0.06169
50	0.04071
51	0.02395
Total	**0.75404**

- **Mean**

  $\mu = np = 60(3/4) = 45$

- **Standard deviation**

  $$\begin{aligned}\sigma &= \sqrt{np(1-p)} \\ &= \sqrt{60(3/4)(1/4)} \\ &= \sqrt{11.25} \approx 3.354101966\end{aligned}$$

- **Mode (k with max pdf)**

  $$\begin{aligned}&\lfloor np + p \rfloor \\ &= \lfloor 60(3/4) + (3/4) \rfloor \\ &= \left\lfloor 45\tfrac{3}{4} \right\rfloor = 45\end{aligned}$$

Mode of a distribution

The *mode* of random variable X is the value k at which the pdf is maximum.

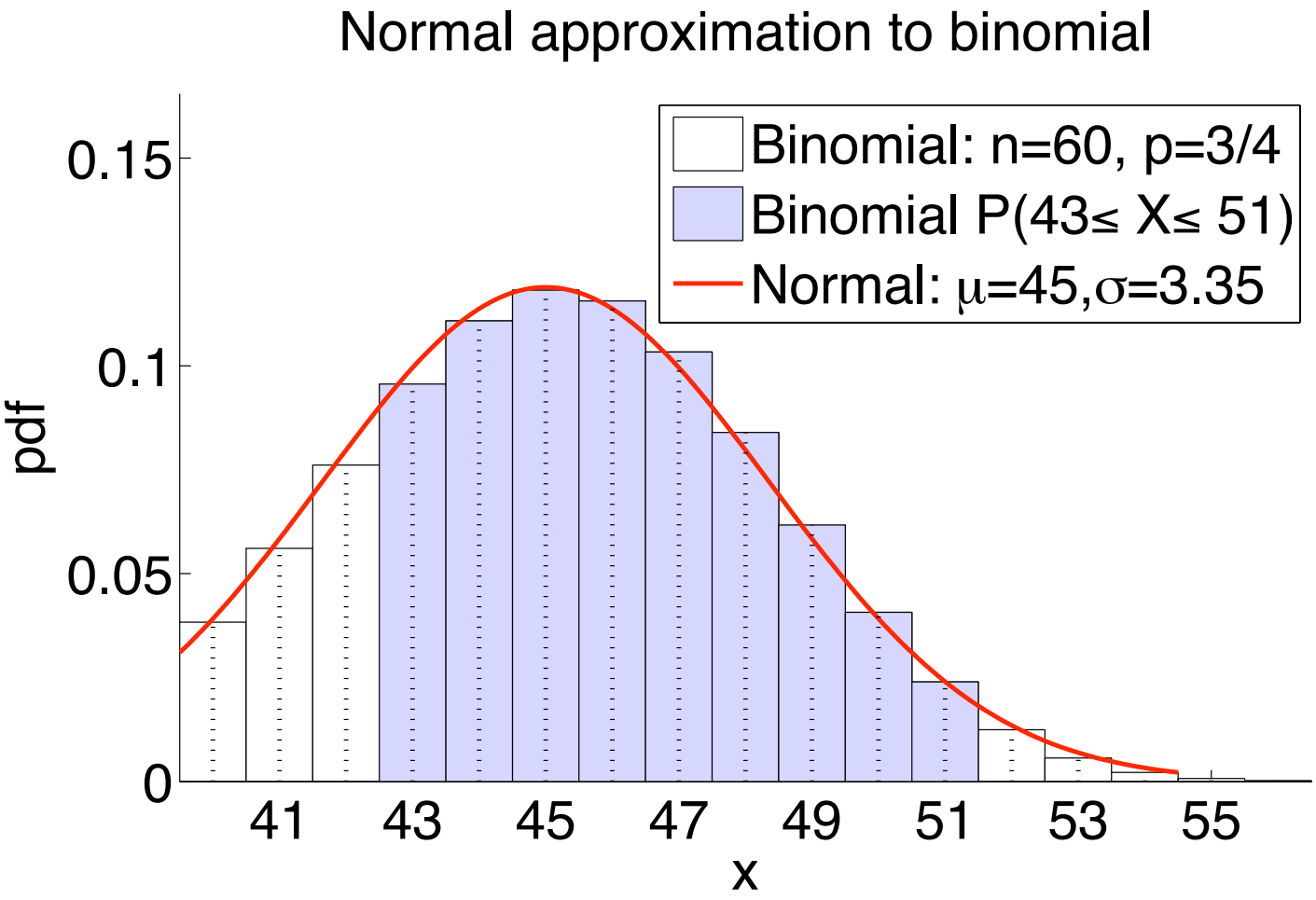
Mode of binomial distribution when $0 < p < 1$

- The mode is $\lfloor (n+1)p \rfloor$.
- *Exception:* If $(n+1)p$ is an integer then $(n+1)p$ and $(n+1)p - 1$ are tied as the mode.
- The mode is within 1 of the mean np.
- When np is an integer, the mode equals the mean.

Binomial and normal distributions

Binomial

k	$P(X = k)$
43	0.09562
44	0.11083
45	0.11822
46	0.11565
47	0.10335
48	0.08397
49	0.06169
50	0.04071
51	0.02395
Total	**0.75404**



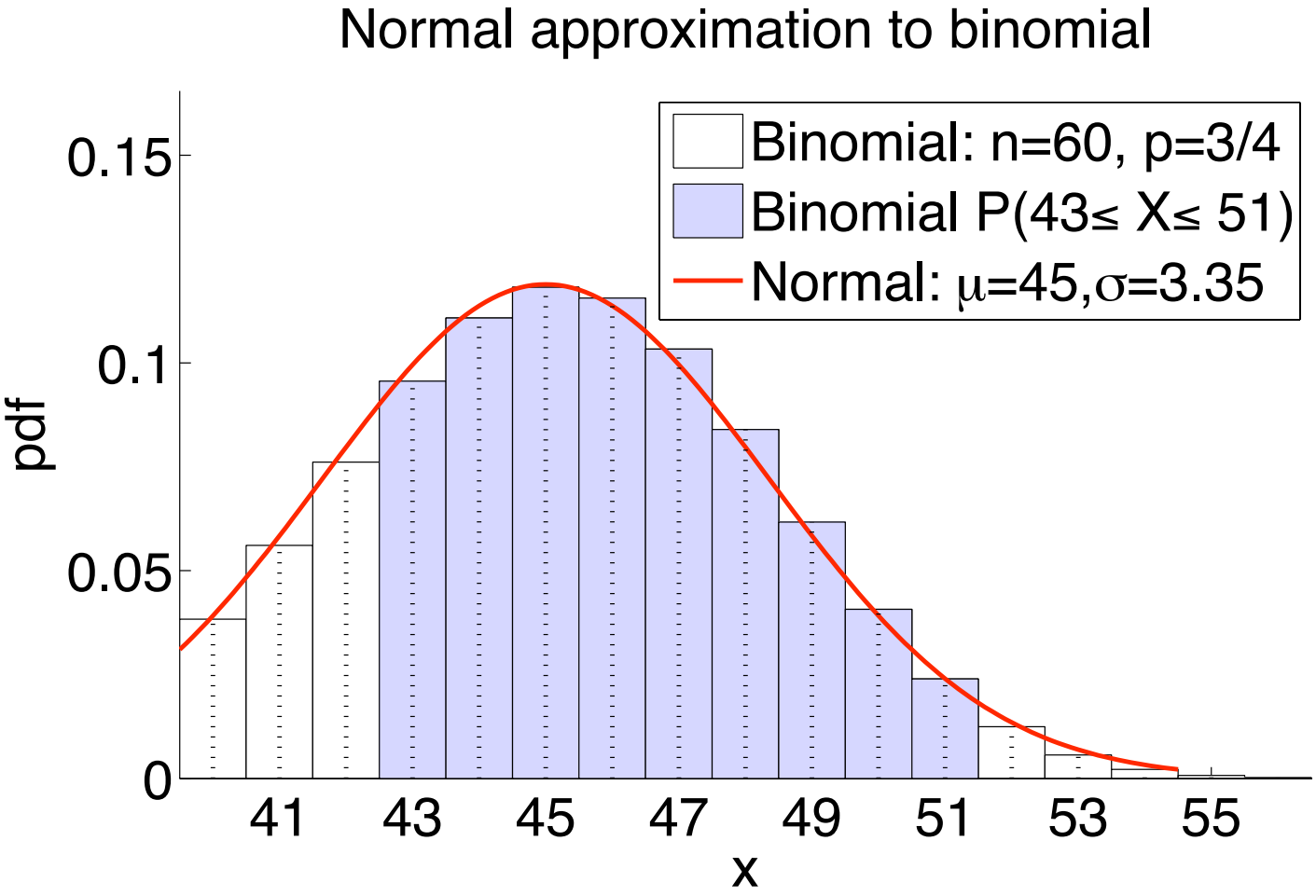
$P(X = k)$ is shown as a rectangle centered above $X = k$:

- Height $P(X = k)$.
- Extent $k \pm 1/2$ gives width 1.
- Area $1 \cdot P(X = k) = P(X = k)$.
- Area of all purple rectangles is $P(43 \leqslant X \leqslant 51)$.

Binomial and normal distributions

Binomial

k	$P(X = k)$
43	0.09562
44	0.11083
45	0.11822
46	0.11565
47	0.10335
48	0.08397
49	0.06169
50	0.04071
51	0.02395
Total	**0.75404**



- The binomial distribution is only defined at the integers, and is very close to the normal distribution shown.
- We will approximate the probability $P(43 \leqslant X \leqslant 51)$ we had above by the corresponding one for the normal distribution.
- Riemann sums in Calculus: area under curve $\approx$ area of rectangles

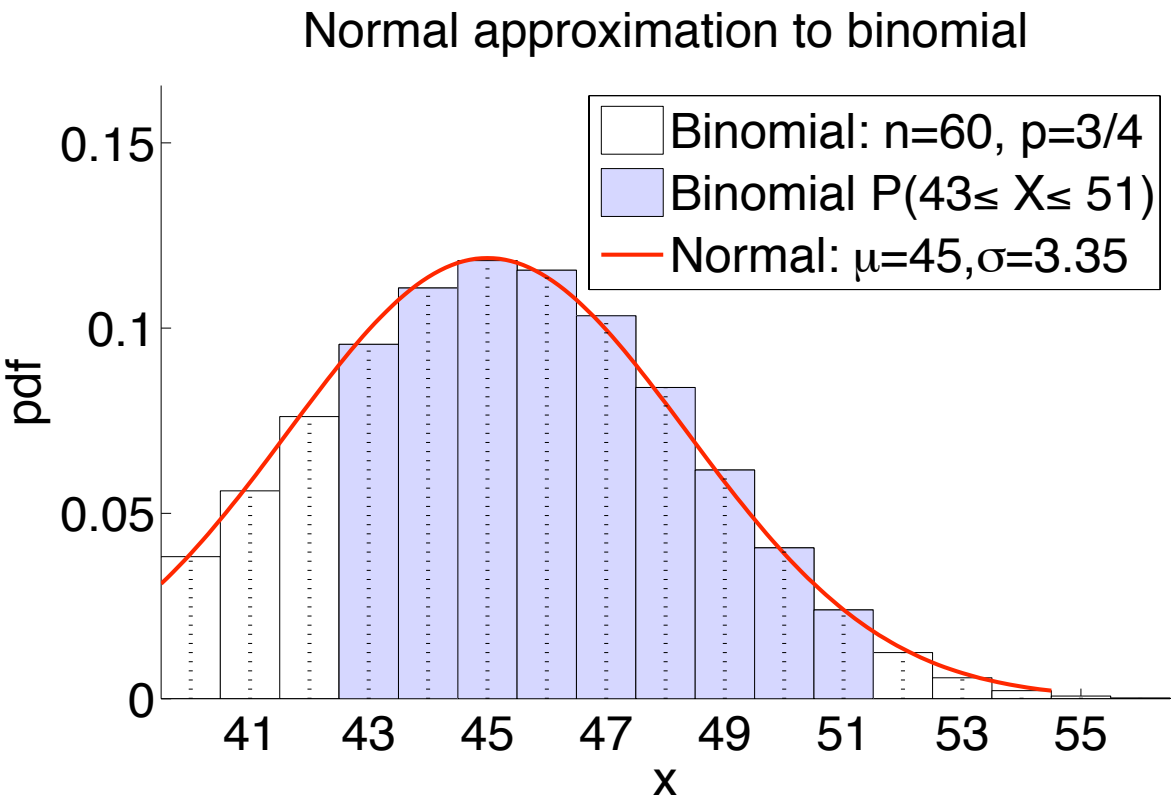
Normal approximation to binomial, step 1

Compute corresponding parameters

- We want to approximate $P(a \leqslant X \leqslant b)$ in a binomial distribution. We'll use $n = 60$, $p = 3/4$ and approximate $P(43 \leqslant X \leqslant 51)$.
- **Determine μ, σ:** $$\mu = np = 60(3/4) = 45$$ $$\sigma = \sqrt{np(1-p)} = \sqrt{11.25} \approx 3.354$$
- The normal distribution with those same values of μ, σ is a good approximation to the binomial distribution *provided* $\mu \pm 3\sigma$ are both between 0 and n.
- **Check:** $$\mu - 3\sigma \approx 45 - 3(3.354) = 34.938$$ $$\mu + 3\sigma \approx 45 + 3(3.354) = 55.062$$ are both between 0 and 60, so we may proceed.
- **Note:** Some applications are more strict and may require $\mu \pm 5\sigma$ or more to be between 0 and n. Since $\mu + 5\sigma \approx 61.771$, this would fail at that level of strictness.

Normal approximation to binomial, step 2

Continuity correction



- The binomial distribution is discrete ($X =$ integers) but the normal distribution is continuous.

- The sum $P(X = 43) + \cdots + P(X = 51)$ has 9 terms, corresponding to the area of the 9 rectangles in the picture.
- The area under the normal distribution curve from $42.5 \leqslant X \leqslant 51.5$ approximates the area of those rectangles.
- Change binomial $P(43 \leqslant X \leqslant 51)$ to normal $P(42.5 \leqslant X \leqslant 51.5)$.

Normal approximation to binomial, step 3

Convert to z-scores

- For random variable X with mean μ and standard deviation σ,
  - The z-score of a value x is $z = \frac{x - E(X)}{\mathrm{SD}(X)} = \frac{x-\mu}{\sigma}$.
  - The random variable Z is $Z = \frac{X - E(X)}{\mathrm{SD}(X)} = \frac{X-\mu}{\sigma}$.
- Convert to z-socres:

$$\begin{aligned} P(42.5 \leqslant X \leqslant 51.5) &= P\left(\frac{42.5-45}{\sqrt{11.25}} \leqslant \frac{X-45}{\sqrt{11.25}} \leqslant \frac{51.5-45}{\sqrt{11.25}}\right) \\ &= P(-.7453559926 \leqslant Z \leqslant 1.937925581) \end{aligned}$$

Normal approximation to binomial, step 4

Use the normal distribution

- We're at $P(43 \leqslant X \leqslant 51) = P(-.7453559926 \leqslant Z \leqslant 1.937925581)$.
- Approximate this by the standard normal distribution cdf:
$$\begin{aligned} &P(-.7453559926 \leqslant Z \leqslant 1.937925581) \\ &\approx \Phi(1.937925581) - \Phi(-.7453559926) \end{aligned}$$
- Look in the standard normal table in the back of the book:
  - It only has z's to two decimal places, so round them: $\Phi(1.94) \approx 0.9738$ and $\Phi(-.75) \approx 0.2266$.
  - So $\Phi(1.937925581) - \Phi(-.7453559926) \approx \boxed{\mathbf{0.7472}}$.
- On a calculator/computer with more digits, it's $\boxed{\mathbf{0.7456555785}}$.
- These are close to the true answer (apart from rounding errors) $P(43 \leqslant X \leqslant 51) = 0.75404$ we got from the binomial distribution.

Estimating fraction of successes instead of number of successes

- What is the value of p in the binomial distribution?
- Estimate it: flip a coin n times and divide the # heads by n.
- Let X = binomial distribution for n flips, probability p of heads.
- Let $\overline{X} = X/n$ be the fraction of flips that are heads.
- $\overline{X}$ is discrete, with possible values $\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$.
- $P(\overline{X} = \frac{k}{n}) = P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } k = 0, 1, \ldots, n; \\ 0 & \text{otherwise.} \end{cases}$
- **Mean** $E(\overline{X}) = E(X/n) = E(X)/n = np/n = p$.
- **Variance** $\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{\mathrm{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$.
- **Standard deviation** $\mathrm{SD}(\overline{X}) = \sqrt{p(1-p)/n}$.

Normal approximation for fraction of successes

- n flips, probability p of heads, $\overline{X}$=observed fraction of heads

Mean	$E(\overline{X}) = p$
Variance	$\mathrm{Var}(\overline{X}) = p(1-p)/n$
Standard deviation	$\mathrm{SD}(\overline{X}) = \sqrt{p(1-p)/n}$

- The Z transformation of $\overline{X}$ is

$$Z = \frac{\overline{X} - E(\overline{X})}{\mathrm{SD}(\overline{X})} = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}}$$

  and value $\overline{X} = \bar{x}$ has z-score $z = \frac{\bar{x}-p}{\sqrt{p(1-p)/n}}$.

- For k heads in n flips,
  - The z-score of $X = k$ is $z_1 = \frac{k-np}{\sqrt{np(1-p)}}$.
  - The z-score of $\overline{X} = k/n$ is $z_2 = \frac{(k/n)-p}{\sqrt{p(1-p)/n}}$.
  - These are equal! Divide the numerator and denominator of z_1 by n to get z_2.

Normal approximation for fraction of successes

- For $n = 60$ flips of a coin with $p = \frac{3}{4}$, we'll estimate $P\left(\frac{43}{60} \leqslant \overline{X} \leqslant \frac{51}{60}\right)$.
- The exact answer equals $P(43 \leqslant X \leqslant 51) \approx 0.75404$.

- **Step 1: Determine mean and SD**
  $E(\overline{X}) = p = .75$
  $\mathrm{SD}(\overline{X}) = \sqrt{p(1-p)/n} = \sqrt{(.75)(.25)/60} = \sqrt{0.003125} \approx 0.05590$
- **Verify approximation is valid: Mean $\pm$ 3 SD between 0 and 1**
  Mean $-$ 3 SD = 0.58229
  Mean + 3 SD = 0.91770
  Both are between 0 and 1.
- **Step 2: Continuity correction**
  $$P\left(\frac{43}{60} \leqslant \overline{X} \leqslant \frac{51}{60}\right) = P\left(\frac{42.5}{60} \leqslant \overline{X} \leqslant \frac{51.5}{60}\right)$$
- **Step 3: z-scores**
- **Step 4: Evaluate using table or calculator**

Normal approximation for fraction of successes

$$
\begin{aligned}
P\left(\tfrac{43}{60} \leqslant \overline{X} \leqslant \tfrac{51}{60}\right) &= P\left(\tfrac{42.5}{60} \leqslant \overline{X} \leqslant \tfrac{51.5}{60}\right) \\
&= P(0.70833 \leqslant \overline{X} \leqslant .85833) \\
&= P\left(\tfrac{0.70833-E(\overline{X})}{\mathrm{SD}(\overline{X})} \leqslant \tfrac{\overline{X}-E(\overline{X})}{\mathrm{SD}(\overline{X})} \leqslant \tfrac{.85833-E(\overline{X})}{\mathrm{SD}(\overline{X})}\right) \\
&= P\left(\tfrac{0.70833-.75}{0.05590} \leqslant Z \leqslant \tfrac{.85833-.75}{0.05590}\right) \\
&= P(-.74535 \leqslant Z \leqslant 1.93792) \\
&= \cdots \approx \boxed{\mathbf{0.7472}} \quad \text{with table in book}
\end{aligned}
$$

or 0.74565 with a calculator/computer

Mean and SD of sums and averages of i.i.d. random variables

- Let $X_1, \ldots, X_n$ be n i.i.d. (independent identically distributed) random variables, each with mean μ and standard deviation σ.
- Let $S_n = X_1 + \cdots + X_n$ be their sum and $\overline{X}_n = (X_1 + \cdots + X_n)/n = S_n/n$ be their average.
  - **Means:**
    Sum: $E(S_n) = E(X_1) + \cdots + E(X_n) = n\,E(X_1) = n\mu$
    Avg: $E(\overline{X}_n) = E(S_n/n) = n\mu/n = \mu$
  - **Variance:**
    Sum: $\mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\,\mathrm{Var}(X_1) = n\sigma^2$
    Avg: $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}(S_n)/n^2 = n\sigma^2/n^2 = \sigma^2/n$
  - **Standard deviation:**
    Sum: $\mathrm{SD}(S_n) = \sigma\sqrt{n}$
    Avg: $\mathrm{SD}(\overline{X}_n) = \sigma/\sqrt{n}$

Terminology for different types of standard deviation

- The *standard deviation* (SD) of a trial (each X_i) is σ
- The *standard error* (SE) of the sum is $\sigma\sqrt{n}$
- The *standard error* (SE) of the average is $\sigma/\sqrt{n}$

Z-scores of sums and averages

	For sum S_n	For average $\overline{X}_n$
Mean:	$E(S_n) = n\mu$	$E(\overline{X}_n) = \mu$
Variance:	$\mathrm{Var}(S_n) = n\sigma^2$	$\mathrm{Var}(\overline{X}_n) = \sigma^2/n$
Standard Deviation:	$\mathrm{SD}(S_n) = \sigma\sqrt{n}$	$\mathrm{SD}(\overline{X}_n) = \sigma/\sqrt{n}$
Z-scores:	$Z = \frac{S_n - E(S_n)}{\mathrm{SD}(S_n)} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$	$Z = \frac{\overline{X}_n - E(\overline{X}_n)}{\mathrm{SD}(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$

Z-scores of sum and average are equal! Divide the numerator and denominator of Z of the sum by n to get Z of the average.

$$Z_{\text{sum}} = \frac{(S_n - n\mu)/n}{(\sigma\sqrt{n})/n} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = Z_{\text{avg}}$$

Theorem (Central Limit Theorem — abbreviated CLT)

For n i.i.d. random variables $X_1, \ldots, X_n$ with sum $S_n = X_1 + \cdots + X_n$ and average $\overline{X}_n = S_n/n$, and any real numbers $a < b$,

$$P\left(a \leqslant \frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant b\right) = P\left(a \leqslant \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant b\right) \approx \Phi(b) - \Phi(a)$$

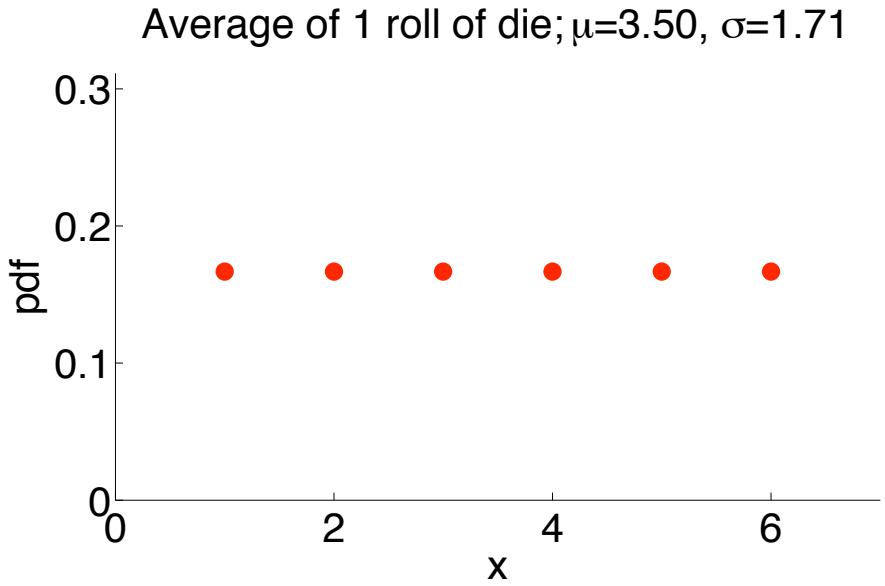
if n is large enough. As $n \to \infty$, the approximation becomes equality.

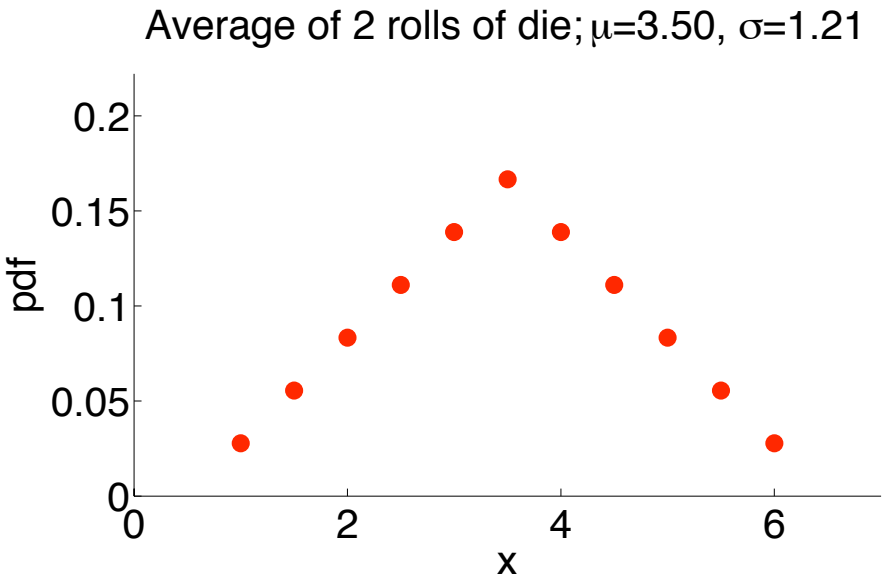
Interpretation of Central Limit Theorem

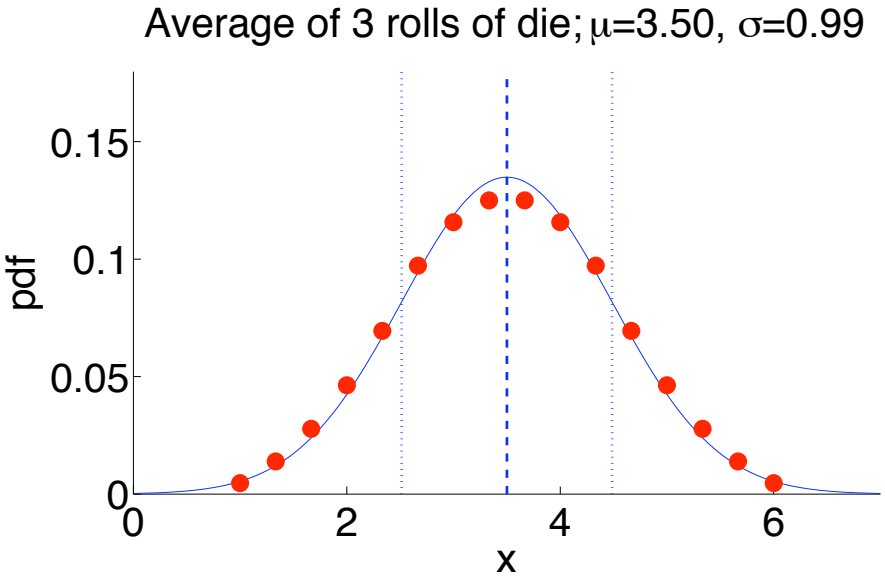
- As n increases, the pdf closely resembles a normal distribution.
- However, the pdf is defined as 0 in-between the red points shown (on upcoming slides), if it's a discrete distribution.
- The cdfs are approximately equal everywhere on the continuum.
- Probabilities of intervals for sums or averages of enough i.i.d. variables can be approximated with the normal distribution.

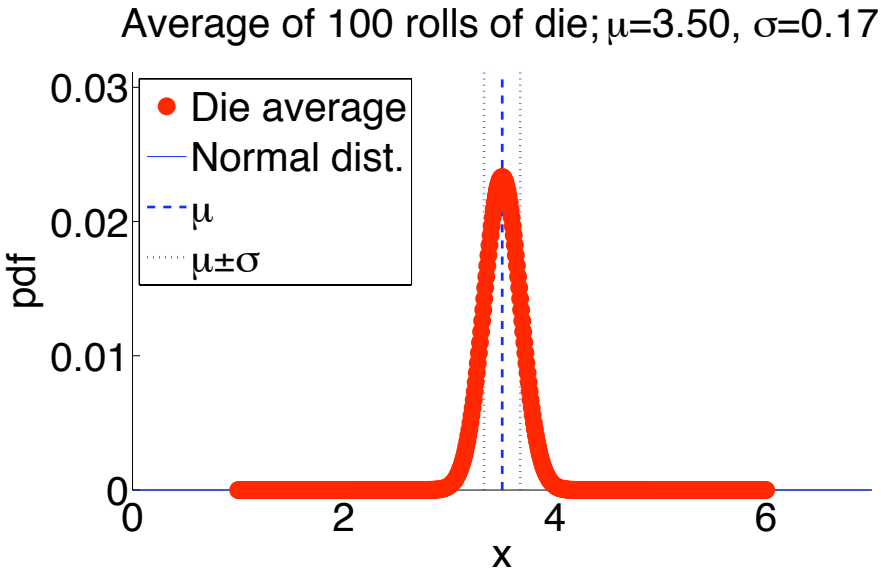
Repeated rolls of a die

One roll: $\mu = 3.5$, $\sigma = \sqrt{35/12} \approx 1.71$









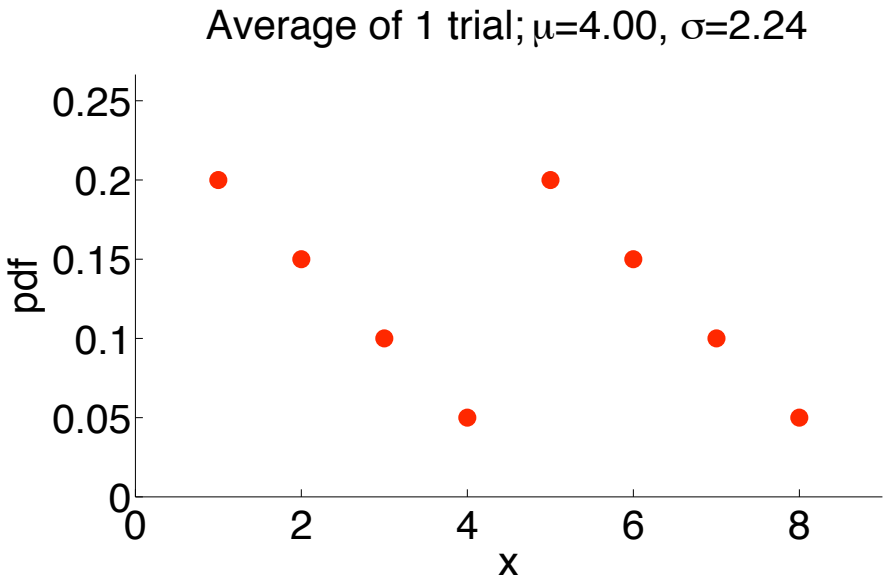
Repeated rolls of a die

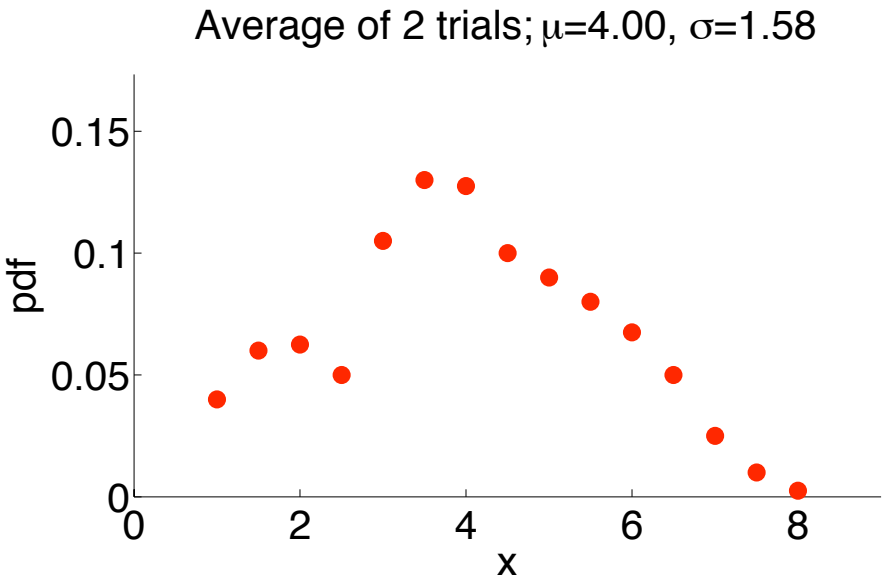
Find n so that at least 95% of the time, the average of n rolls of a die is between 3 and 4.

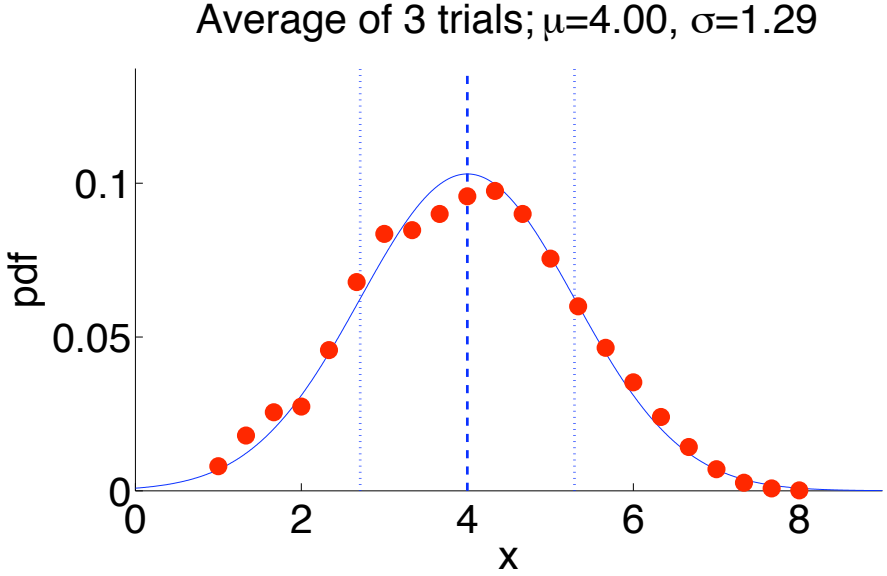
- $P(3 \leqslant \overline{X} \leqslant 4) = P\left(\frac{3-\mu}{\sigma/\sqrt{n}} \leqslant \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leqslant \frac{4-\mu}{\sigma/\sqrt{n}}\right)$
- Plug in $\mu = 3.5$ and $\sigma = \sqrt{35/12}$.
- $P(3 \leqslant \overline{X} \leqslant 4) = P\left(-\frac{1/2}{\sqrt{35/(12n)}} \leqslant Z \leqslant \frac{1/2}{\sqrt{35/(12n)}}\right)$
- Recall the center 95% of the area on the standard normal curve is between $z = \pm 1.96$.
- $\frac{1/2}{\sqrt{35/(12n)}} \geqslant 1.96 \quad \Rightarrow \quad n \geqslant (1.96)^2 \frac{35/12}{(1/2)^2} \approx 44.81$
- n is an integer so $\boxed{n \geq 45}$

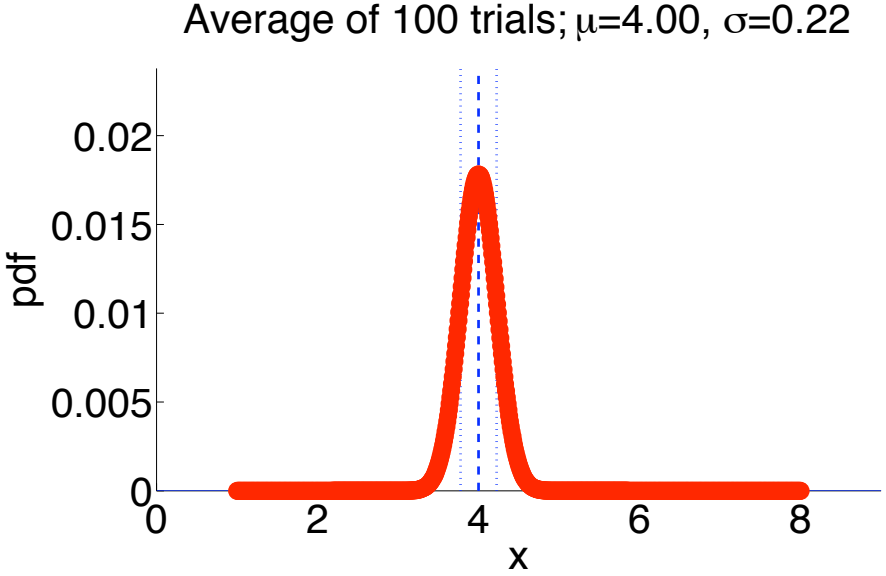
“Sawtooth” distribution (made up as demo)

One trial: $\mu = 4$, $\sigma \approx 2.24$









Binomial distribution (n, p)

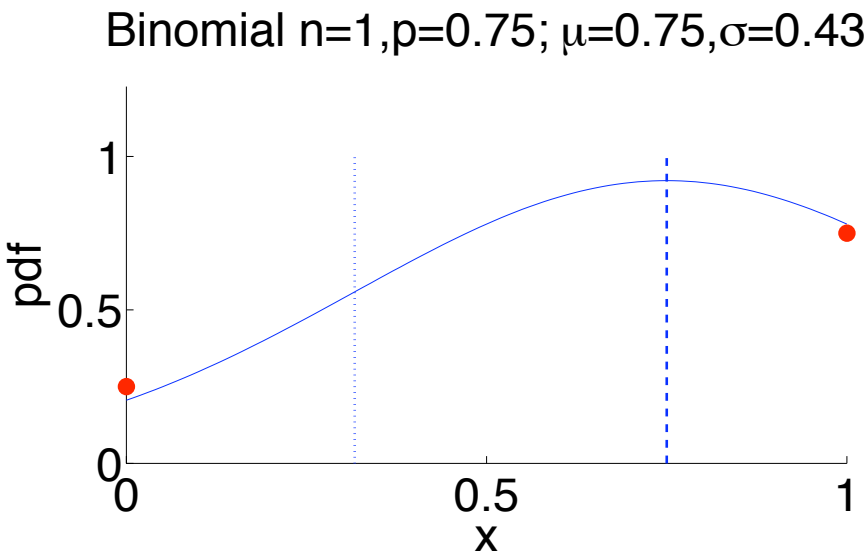
- A *Bernoulli trial* is to flip a coin once and count the number of heads,

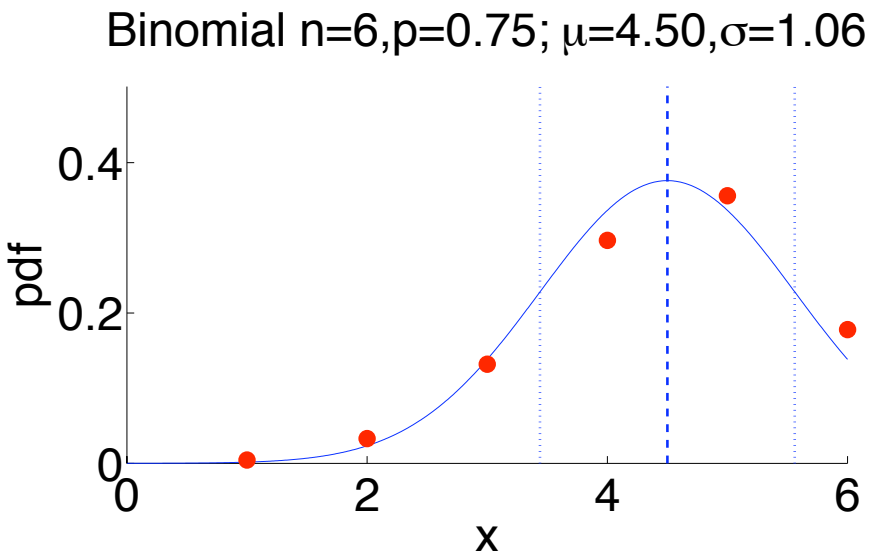
$$X_1 = \begin{cases} 1 & \text{probability } p; \\ 0 & \text{probability } 1-p. \end{cases}$$

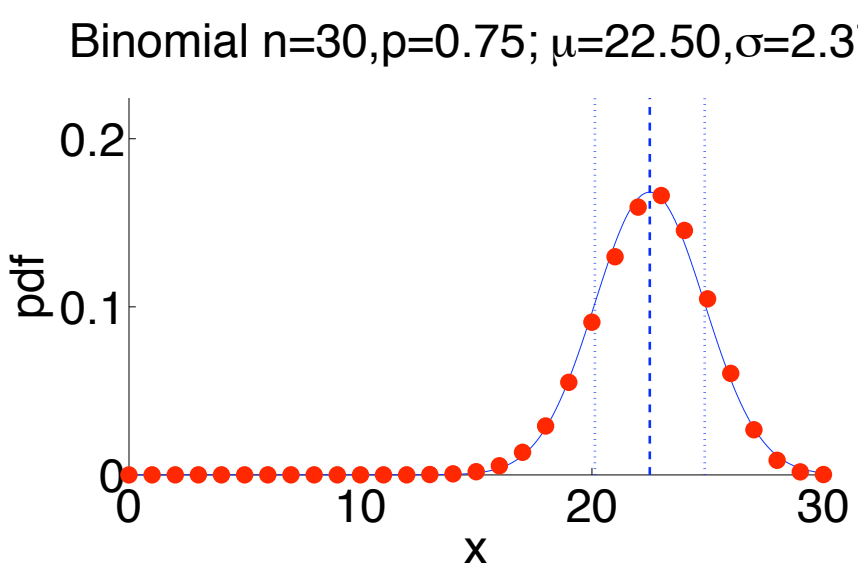
  Mean $E(X_1) = p$, standard deviation $\mathrm{SD}(X_1) = \sqrt{p(1-p)}$.
- The binomial distribution is the sum of n i.i.d. Bernoulli trials. Mean $\mu = np$, standard deviation $\sigma = \sqrt{np(1-p)}$.
- The binomial distribution is approximated pretty well by the normal distribution when $\mu \pm 3\sigma$ are between 0 and n.

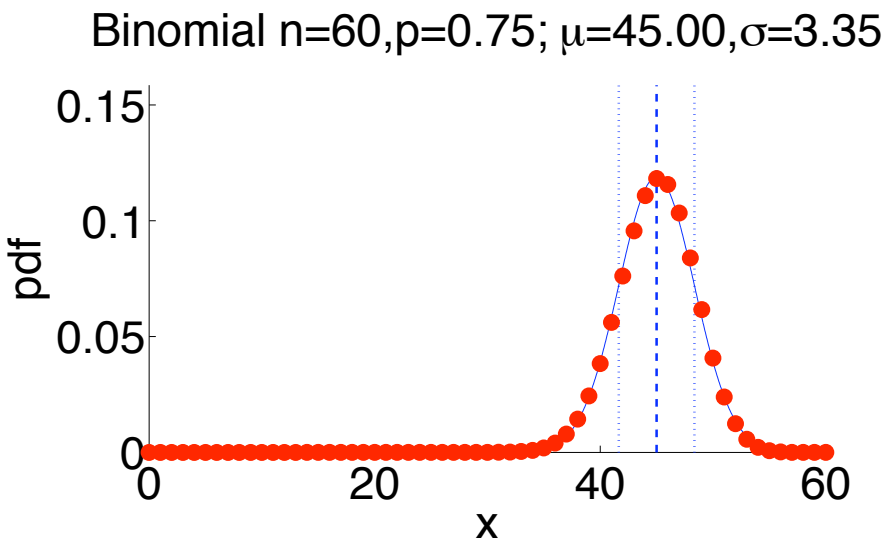
Binomial distribution (n, p)

One flip: $\mu = p = .75$, $\sigma = \sqrt{p(1-p)} = \sqrt{.1875} \approx 0.4330$







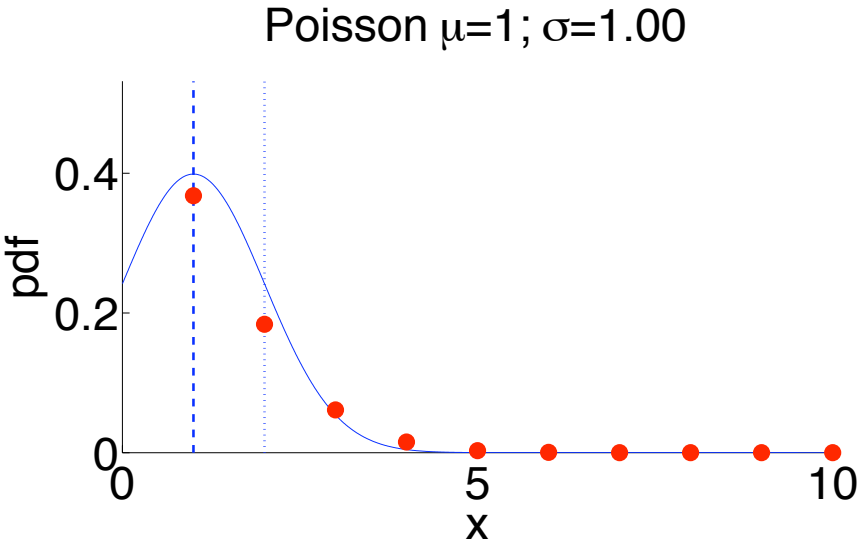


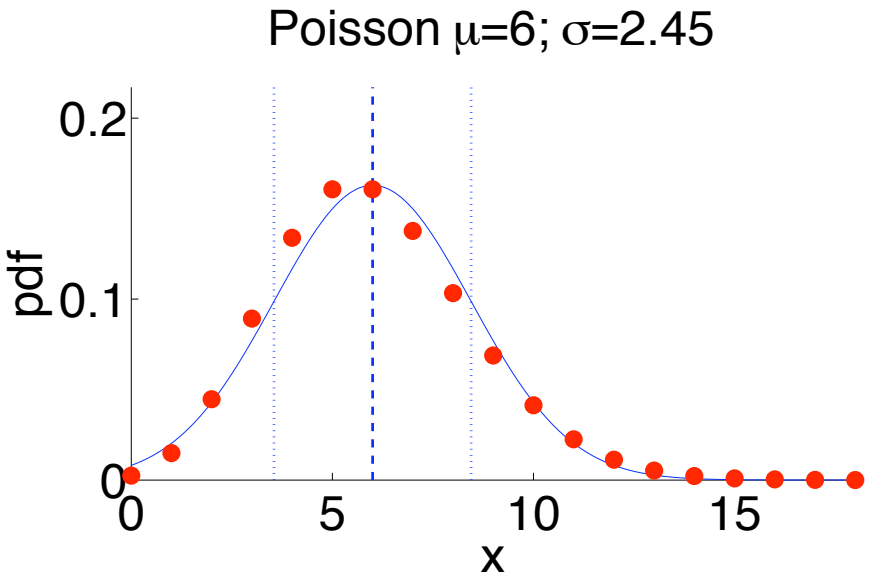
Poisson distribution (μ or $\mu = \lambda\, d$)

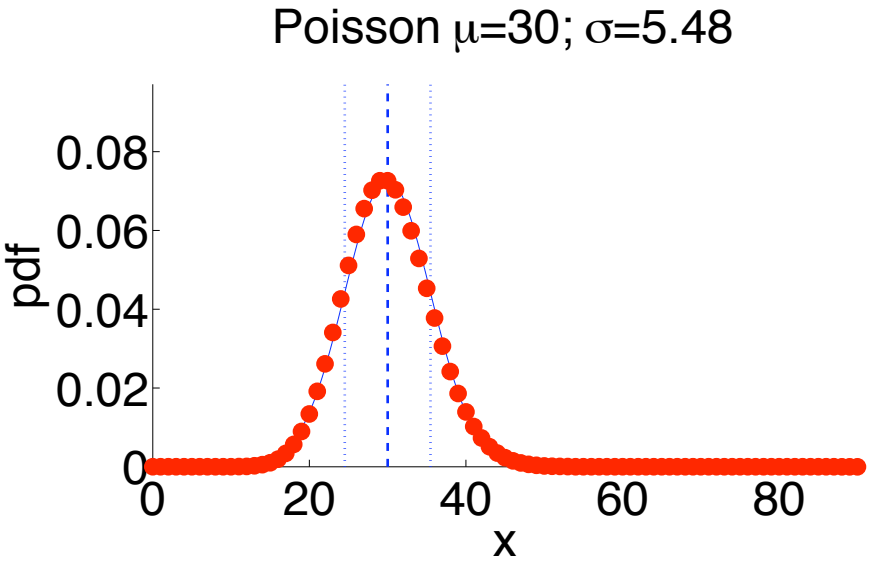
- **Mean:** μ (same as the Poisson parameter)
  Standard deviation: $\sigma = \sqrt{\mu}$.
- It is approximated pretty well by the normal distribution when $\mu \geqslant 5$.
- The reason the Central Limit Theorem applies is that a Poisson distribution with parameter μ equals the sum of n i.i.d. Poissons with parameter μ/n.

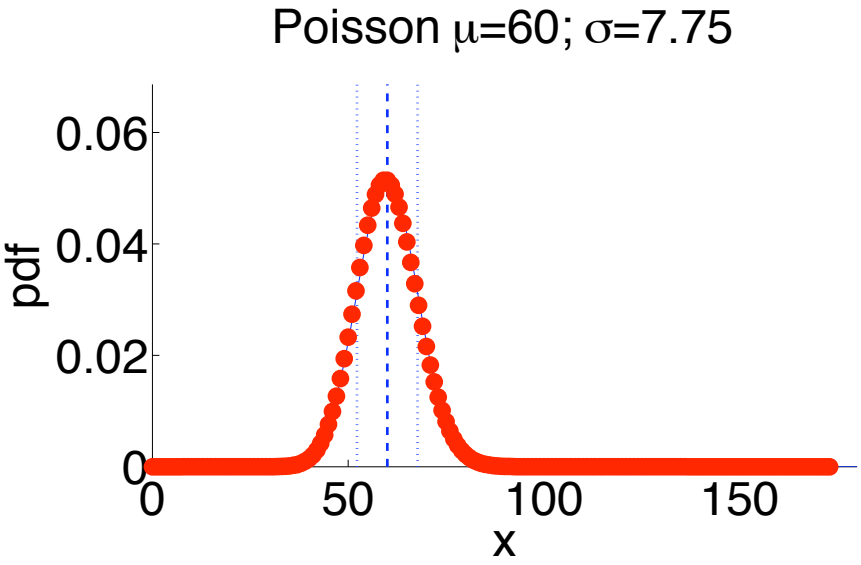
- The Poisson distribution has infinite range $x = 0, 1, 2, \ldots$ and the normal distribution has infinite range $-\infty < x < \infty$ (reals). Both are truncated in the plots.

Poisson distribution (μ)









Geometric and negative binomial distributions

Geometric distribution (p)

- X is the number of flips until the first heads,

$$p_X(x) = \begin{cases} (1-p)^{x-1}p & \text{if } x = 1, 2, 3, \ldots; \\ 0 & \text{otherwise.} \end{cases}$$

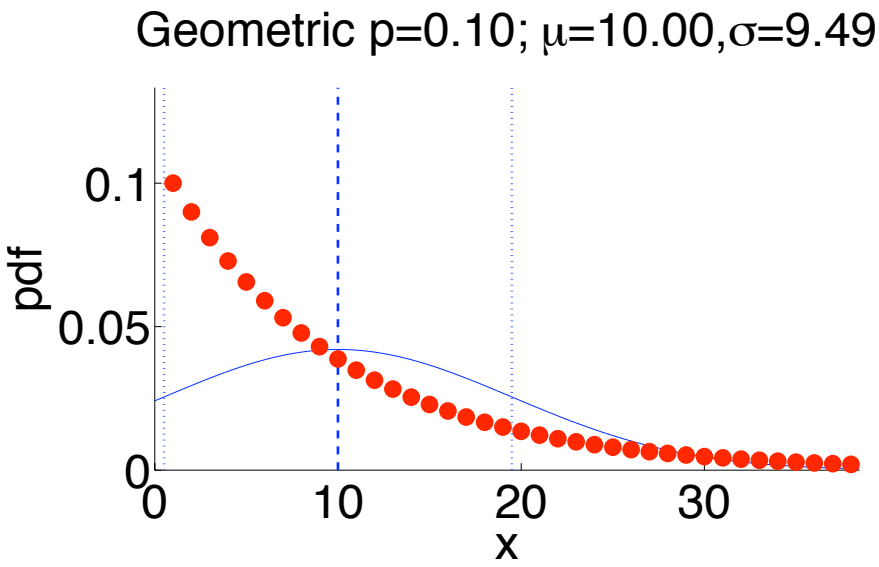
- The pdf plot doesn't resemble the normal distribution at all.
- **Mean:** $\mu = 1/p$ **Standard deviation:** $\sigma = \sqrt{1-p}/p$

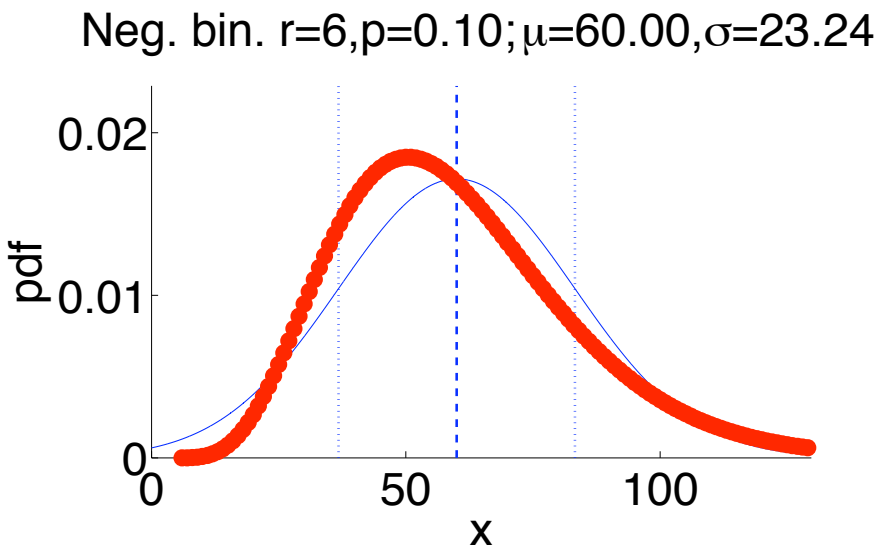
Negative binomial distribution (r, p)

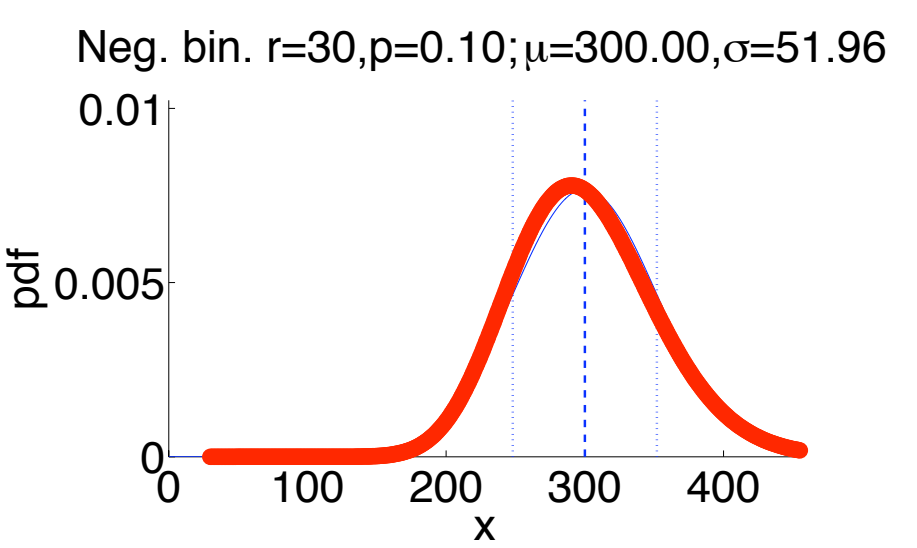
- $r = 1$ is same as geometric distribution.
- $r > 2$: The pdf has a "bell"-like shape, but is not close to the normal distribution unless r is very large.
- **Mean:** $\mu = r/p$ **Standard deviation:** $\sigma = \sqrt{r(1-p)}/p$

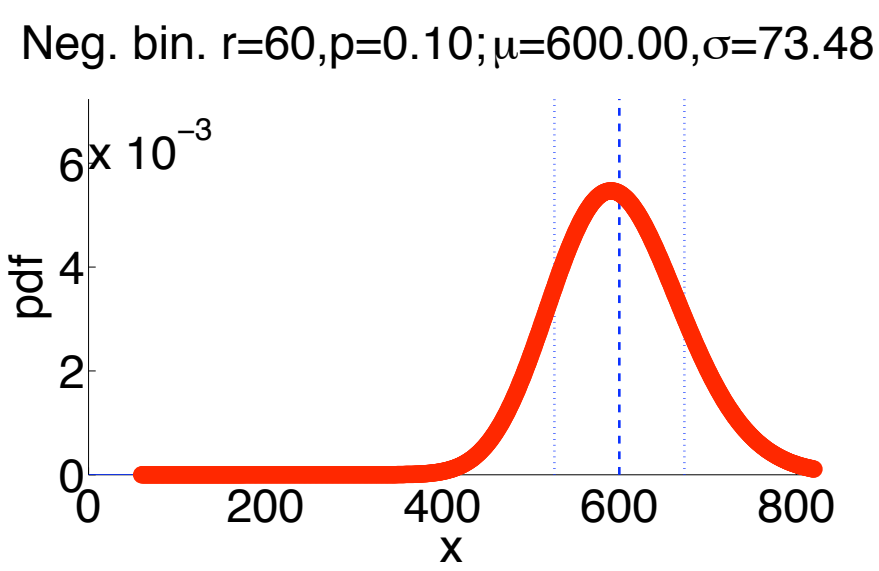
Geometric and negative binomial distributions

Heads with probability $p = .1$









Exponential and gamma distributions

Exponential distribution (λ)

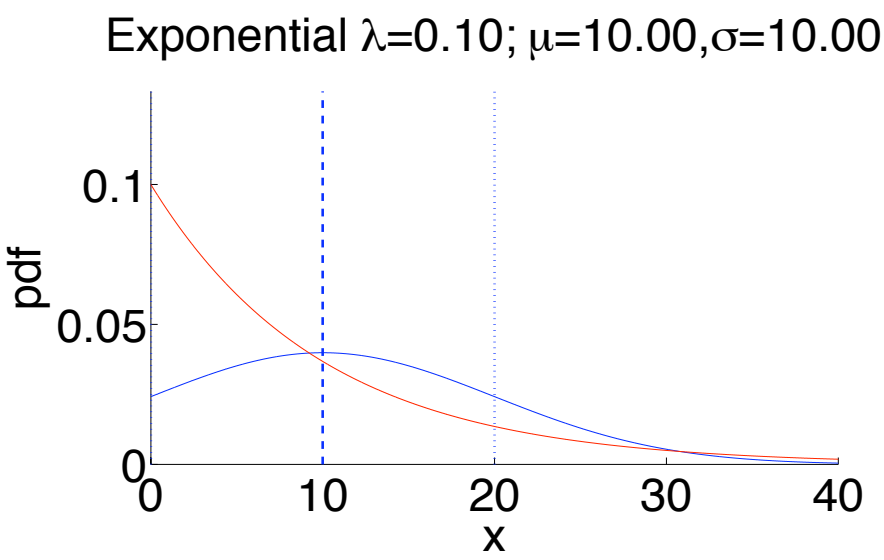
- The exponential distribution doesn't resemble the normal distribution at all.
- **Mean:** $\mu = 1/\lambda$ **Standard deviation:** $\sigma = 1/\lambda$

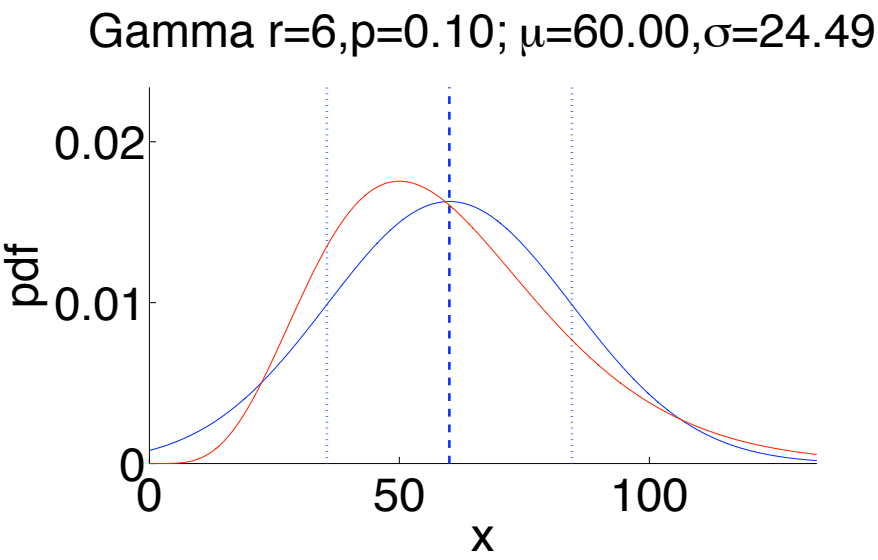
Gamma distribution (r, λ)

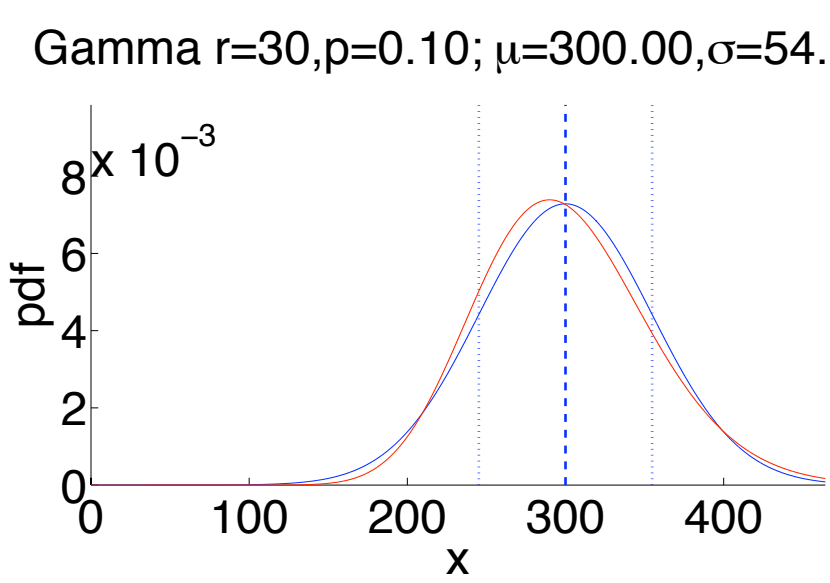
- The gamma distribution for $r = 1$ is the exponential distribution.
- The gamma distribution for $r > 1$ does have a "bell"-like shape, but it is not close to the normal distribution until r is very large.
- There is a generalization to allow r to be real numbers, not just integers.
- **Mean:** $\mu = r/\lambda$ **Standard deviation:** $\sigma = \sqrt{r}/\lambda$

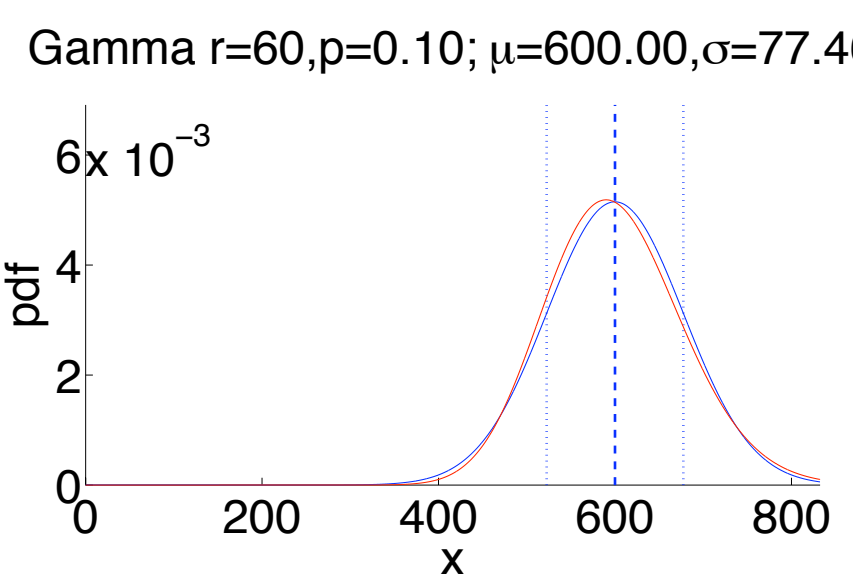
Exponential and gamma distributions

Rate $\lambda = .1$









Geometric/Negative binomial vs. Exponential/Gamma

- $p = \lambda$ gives same means for geometric and exponential.
- $p = 1 - e^{-\lambda}$ gives same exponential decay rate for both geometric and exponential distributions.
- $1 - e^{-\lambda} \approx \lambda$ when λ is small.
- This corespondence carries over to the gamma and negative binomial distributions.

Geometric/negative binomial vs. Exponential/gamma

This is for $p = .1$ vs. $\lambda = .1$; a better fit for $\lambda = .1$ would be $p = 1 - e^{-\lambda} \approx 0.095$