Math 425
Intro to Probability
Lecture 22

Kenneth Harris
kaharri@umich.edu

Department of Mathematics
University of Michigan

March 9, 2009

Example: Grading on a curve

Example

Final exams at Podunk U. are constructed so that the distribution of scores is approximately normally distributed, with parameters μ (the average score) and σ (the standard deviation from the average). Letter grades are then assigned according to the following chart:

Test Score	Grade
$\mu + \sigma < x$	A
$\mu < x < \mu + \sigma$	B
$\mu - \sigma < x < \mu$	C
$\mu - 2\sigma < x < \mu - \sigma$	D
$x < \mu - 2\sigma$	F

☞ This system of assigning letter grades is called “grading on the curve".

Example: Grading on a curve

☞ Let X be a normally distributed r.v. with parameters μ and σ.
By standardization: $Z = \frac{X-\mu}{\sigma}$.

$$\begin{aligned}
\mathbf{P}\{\mu + \sigma < X\} &= \mathbf{P}\{1 < \frac{X-\mu}{\sigma}\} = 1 - \Phi(1) \approx 0.1587 \\
\mathbf{P}\{\mu < X < \mu + \sigma\} &= \mathbf{P}\{0 < \frac{X-\mu}{\sigma} < 1\} = \Phi(1) - \Phi(0) \approx 0.3413 \\
\mathbf{P}\{\mu - \sigma < X < \mu\} &= \mathbf{P}\{-1 < \frac{X-\mu}{\sigma} < 0\} \\
&= \Phi(0) - \Phi(-1) = \Phi(0) + \Phi(1) - 1 \approx 0.3413 \\
\mathbf{P}\{\mu - 2\sigma < X < \mu - \sigma\} &= \mathbf{P}\{-2 < \frac{X-\mu}{\sigma} < -1\} \\
&= \Phi(-2) - \Phi(-1) = \Phi(1) - \Phi(2) \approx 0.1359 \\
\mathbf{P}\{X < \mu - 2\sigma\} &= \mathbf{P}\{\frac{X-\mu}{\sigma} < -2\} = 1 - \Phi(2) = 0.0228
\end{aligned}$$

The probabilities can be computed from a table for the standard normal curve.

Example: Grading on a curve

☞ The final exam letter grades at Podunk U. are distributed as follows:

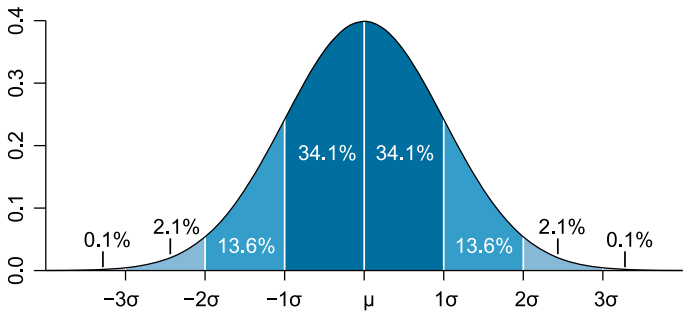
Test Score	Grade	Percentage
$\mu + \sigma < x$	A	$\sim 16\%$
$\mu < x < \mu + \sigma$	B	$\sim 34\%$
$\mu - \sigma < x < \mu$	C	$\sim 34\%$
$\mu - 2\sigma < x < \mu - \sigma$	D	$\sim 14\%$
$x < \mu - 2\sigma$	F	$\sim 2\%$

Standard Deviations

☞

Std deviation	Probability
$\pm\sigma$	0.6827
$\pm 2\sigma$	0.9545
$\pm 3\sigma$	0.9973

http://upload.wikimedia.org/wikipedia/commons/8/8c/Standard_deviation_diagram.svg



Discrete Approximations

☞ In many practical situations, complicated and unwieldy distributions can be usefully replaced by simpler approximations.

- The hypergeometric distribution can be approximated by the binomial distribution (when the population is large compared to the sample size).
- The binomial distribution can be approximated by a Poisson distribution (when successes are rare – very large population and very small probability).

Continuous Approximations

☞ The common discrete distributions also have continuous approximations, which are very good in the limit.

Discrete Distr.	Continuous Distr.	Section
uniform	uniform	5.3
geometric	exponential	5.5
binomial	normal	5.4
neg. binomial	Gamma	5.6

☞ In the last two cases, the continuous distribution is considerably easier to compute and can be very accurate.

Uniform Approximation

Example

Let V_n be a discrete uniform distribution on $\{1, 2, \ldots, n\}$ and U the continuous uniform distribution on $[0, 1]$ with cumulative distribution F_U.

Then for all reals a and b:

$$\lim_{n\to\infty} \mathbf{P}\left\{a \le \frac{V_n}{n} \le b\right\} = F_U(b) - F_U(a).$$

Discrete Uniform Distribution

Definition. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer no bigger than x.

☞ Let V_n be the discrete uniform distribution on $\{1, 2, \ldots, n\}$.
The mass function for $\frac{V_n}{n}$ is

$$p_{V_n/n}(a) = \begin{cases} \frac{1}{n} & \text{if } a \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\} \\ 0 & \text{otherwise.} \end{cases}$$

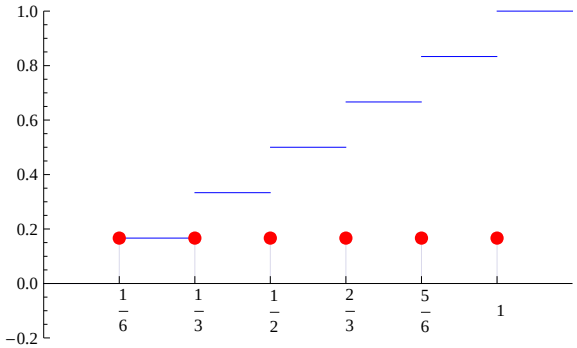
The cumulative distribution is

$$\mathbf{P}\{\frac{V_n}{n} \leq a\} = \begin{cases} 0 & \text{if } a < 0 \\ \frac{\lfloor an \rfloor}{n} & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } 1 < a \end{cases}$$

Note: if $\frac{k}{n} \leq a < \frac{k+1}{n}$ then $\frac{\lfloor an \rfloor}{n} = \frac{k}{n}$.

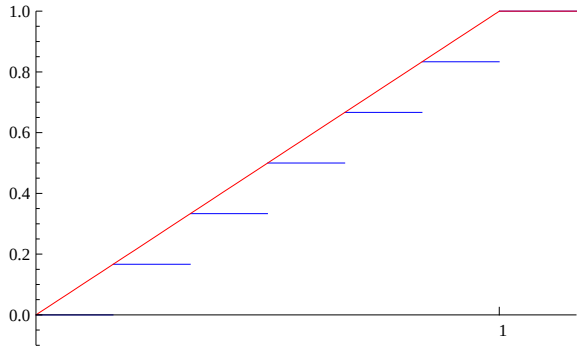
Discrete Mass and Distribution for $n = 6$

☞ mass function and cumulative distribution for $\frac{V_6}{6}$.



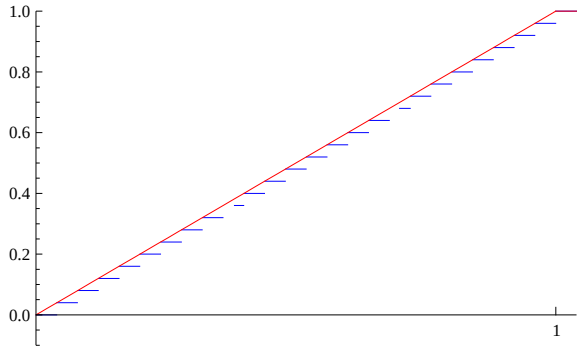
Comparison for $n = 6$

☞ Discrete uniform distribution $\mathbf{P}\{\frac{V_6}{6} \leq a\}$ versus continuous uniform distribution $\mathbf{P}\{U \leq a\}$ on $[0, 1]$.



Comparison for $n = 25$

☞ Discrete uniform distribution $\mathbf{P}\{\frac{V_{25}}{25} \leq a\}$ versus continuous uniform distribution $\mathbf{P}\{U \leq a\}$ on $[0, 1]$.



Uniform Approximation

☞ Let X be a discrete uniform r.v. on $\{1, 2, \ldots, n\}$. Then

$$\mathbf{P}\{\frac{V_n}{n} \le a\} = \begin{cases} 0 & \text{if } a < 0 \\ \frac{\lfloor an \rfloor}{n} & \text{if } 0 \le a \le 1 \\ 1 & \text{if } 1 < a. \end{cases}$$

Note: if $\frac{k}{n} \le a < \frac{k+1}{n}$ then $\frac{\lfloor an \rfloor}{n} = \frac{k}{n}$.

Let U be the continuous uniform r.v. on $[0, 1]$. Then

$$\mathbf{P}\{U \le a\} = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } 0 \le a \le 1 \\ 1 & \text{if } 1 < a. \end{cases}$$

So, for every real number a

$$\left|\mathbf{P}\{\frac{V_n}{n} \le a\} - \mathbf{P}\{U \le a\}\right| \le \frac{1}{n}.$$

Central Limit Theorem

☞ The following is a special case of the Central Limit Theorem, which we will prove it in Chapter 8. (See Ross, p. 225.)

Theorem (DeMoivre-Laplace Limit Theorem)

Let the number of successes in n Bernoulli trials be B_n. So, B_n is a binomial r.v. with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

For any $a < b$ and integer $0 \le k \le n$,

$$\lim_{n\to\infty} \boldsymbol{P}\{a \le \frac{B_n - \mu}{\sigma} \le b\} = \Phi(b) - \Phi(a)$$

$$\boldsymbol{P}\{B_n = k\} \approx \frac{1}{\sigma}\phi(\frac{k-\mu}{\sigma})$$

where Φ is the cumulative distribution and ϕ is the density for the standard normal r.v. Z.

Example

Example

Consider $n = 10^6$ Bernoulli trials with success probability $p = 0.5$. What is the probability that the number of successes lie between $a = 500,000$ and $b = 501,000$?

$$\sum_{k=a}^{b} \binom{10^6}{k} (0.5)^{10^6}.$$

☞ Let X be the binomial random variable. The mean and variance is

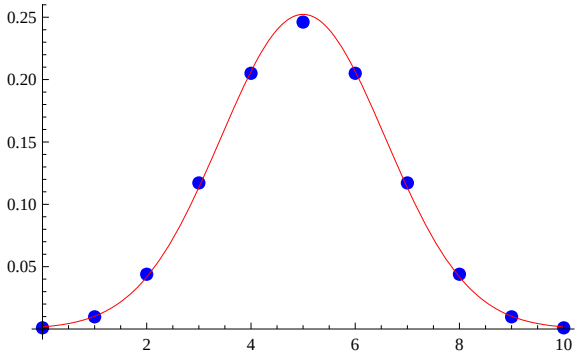
$$\mu = np = 500,000 \qquad \sigma = \sqrt{np(1-p)} = 500.$$

When standardized, X is approximated by Z:

$$\begin{aligned} \mathbf{P}\{a \le X \le b\} &= \mathbf{P}\{0 \le \frac{X-\mu}{\sigma} \le 2\} \\ &\approx \Phi(2) - \Phi(0) \\ &\approx 0.997 - 0.5 = 0.497. \end{aligned}$$

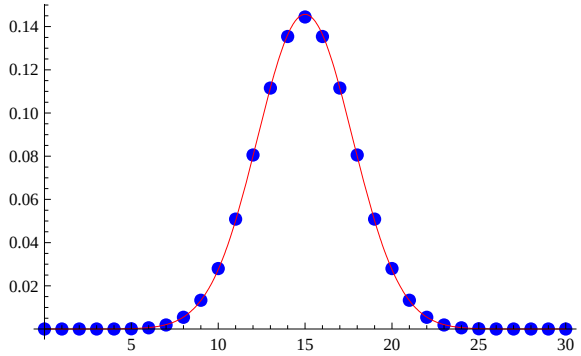
Binomial versus normal

☞ Binomial distribution for $n = 10, p = 0.5$ versus
Normal distribution for $\mu = 5, \sigma^2 = 2.5$.



Binomial versus normal

☞ Binomial distribution for $n = 30, p = 0.5$ versus
Normal distribution for $\mu = 15, \sigma^2 = 7.5$.

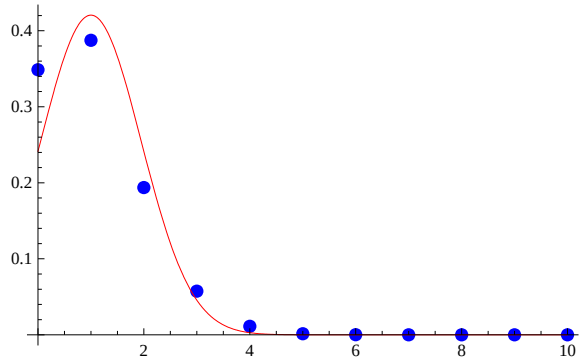


Binomial versus normal

☞ Binomial distribution B_{10} for $n = 10, p = 0.1$ versus
Normal distribution n for $\mu = 1, \sigma^2 = 0.9$.

$\mathbf{P}\{B_{10} = 0\} \approx 0.349$ $\quad \frac{1}{\sigma}\phi(\frac{0-\mu}{\sigma}) \approx 0.241$

$\mathbf{P}\{B_{10} = 1\} \approx 0.387$ $\quad \frac{1}{\sigma}\phi(\frac{1-\mu}{\sigma}) \approx 0.421$



Normal Approximation

☞ The normal approximation is better

- the larger n is,
- the nearer p is to $1 - p$
- the nearer k is to np

☞ The normal approximation is worse

- the smaller n is,
- the smaller p (or $1 - p$) is
- the further k is to np

Continuity Correction

☞ We can improve our approximation of a binomial r.v. X by noting that the discrete binomial random variable takes only integer values.

Continuity correction adjusts by $\mp\frac{1}{2}$, instead of taking endpoints at integer values: when k, m are integers

$$\begin{aligned}\mathbf{P}\{k \le X \le m\} &= \mathbf{P}\{k - \frac{1}{2} \le X \le m + \frac{1}{2}\} \\ &\approx \Phi\left(\frac{m + \frac{1}{2} - \mu}{\sigma}\right) - \Phi\left(\frac{k - \frac{1}{2} - \mu}{\sigma}\right)\end{aligned}$$

$$\begin{aligned}\mathbf{P}\{X = k\} &= \mathbf{P}\{k - \frac{1}{2} \le X \le k + \frac{1}{2}\} \\ &\approx \Phi\left(\frac{k + \frac{1}{2} - \mu}{\sigma}\right) - \Phi\left(\frac{k - \frac{1}{2} - \mu}{\sigma}\right)\end{aligned}$$

Example: How many heads

Example

What is the probability that heads turns up at least 495 times but at most 510 times, when a fair coin tossed 1000 times?

☞ Let X be the r.v. counting heads. Approximate using the normal distribution with $\mu = 500$ and $\sigma = \sqrt{250}$.

$$\begin{aligned}
\mathbf{P}\{494.5 \le X \le 510.5\} &= \mathbf{P}\{\frac{-5.5}{\sqrt{250}} \le \frac{X-500}{\sqrt{250}} \le \frac{10.5}{\sqrt{250}}\} \\
&= \mathbf{P}\{-0.35 \le Z \le 0.66\} \\
&= \Phi(0.66) - \Phi(-0.35) = \Phi(0.66) + \Phi(0.35) - 1 \\
&\approx 0.7454 + 0.6368 - 1 = 0.3822
\end{aligned}$$

Example: Fanatical Gambler

Example

In the course of a year a fanatical gambler makes 10,000 fair wagers. (That is, winning and losing is equally likely.) The gambler wins 4850 of these and loses the rest. Are the Fates against him?

☞ Let X record the number of wins (a binomial r.v.), with $\mu = 5000$ and $\sigma = 50$. Using continuity correction:

$$\begin{aligned}
\mathbf{P}\{X \le 4850.5\} &= \mathbf{P}\{\frac{X-5000}{50} \le -2.99\} \\
&\approx \Phi(-2.99) = 1 - \Phi(2.99) \\
&\approx 1 - 0.9986 \approx 0.0014.
\end{aligned}$$

The fates were most certainly against our gambler.

The successes are -3σ from the mean, a very unlikely outcome.

Example: A cheat or not?

Example

You suspect Hugo the Hustler of using crooked dice (more sixes appear than ought to). Before you accuse you want to be certain. Your plan is to test one of his die by 180 throws, expecting $180 \times \frac{1}{6} = 30$ sixes. You adopt the following rule: if the number of sixes is between 25 and 35 (inclusive) than you will accept it as fair. Otherwise, you will discuss the matter with Enrique the Enforcer to bring justice to the matter.

How sure can you be about this allegation, should the die appear in the crooked region?

Example – continued

☞ Let X be the random variable counting sixes. The probability a fair die lies in in the crooked region is

$$1 - \mathbf{P}\{25 \le X \le 35\} = 1 - \sum_{k=25}^{35} \binom{180}{k} (\frac{1}{6})^k (\frac{5}{6})^{180-k}$$

Even better is to use the normal approximation (with continuity correction):

$$\begin{aligned}
1 - \mathbf{P}\{24.5 \le X \le 35.5\} &= 1 - \mathbf{P}\{-1.1 \le \frac{X-30}{5} \le -1.1\} \\
&\approx 1 - \big(\Phi(1.1) - \Phi(-1.1)\big) = 2 - 2\Phi(1.1) \\
&\approx 2 - 2(0.8643) = 0.2714.
\end{aligned}$$

☞ Hugo has friends, so you might want more a more reliable test.

Example – continued

☞ What if you expand the range of accepable die results to allow the number of sixes to be between 20 and 40 (inclusive).

The normal approximation for the fair die (with continuity correction) is

$$\begin{aligned}\mathbf{P}\{19.5 \le X \le 40.5\} = &= \mathbf{P}\{-2.1 \le \frac{X-30}{5} \le 2.1\} \\ &\approx 1 - \big(\Phi(2.1) - \Phi(-2.1)\big) = 2 - 2\Phi(2.1) \\ &\approx 2 - 2(0.9821) = 0.0358.\end{aligned}$$

☞ This is significantly better. If the die should fail this test, you can take you beef to Enrique.

Example: Too close to call

Example

The UofM is considering a new fight song, one a little less pompous. The administration has contracted with Acme Surveys to do a study of the question. If a majority of the community supports a change, then it will commission a new song.

Suppose 50.5% of the community believes the fight song should NOT change. Acme Surveys will choose n people at random for their sample. How big must n be to ensure with 95% certainty that over half those surveyed want to keep Hail to the Victors?

Example – continued

☞ Let X_n be the random variable which counts those supporting the status quo out of n people in the survey. So, X_n is a hypergeometric random variable.

☞ However, since the University community is large compared to the sample size, X_n is well approximated by a binomial random variable with parameters n and $p = 0.505$.

Example – continued

☞ We approximate X_n with a normal distribution using parameters

$$\mu = 0.505n \qquad \sigma = \sqrt{n(0.505)(0.495)}$$

By standardizing

$$\begin{aligned}\mathbf{P}\{0.5n < X_n\} &= \mathbf{P}\{\frac{0.5n-\mu}{\sigma} < \frac{X_n-\mu}{\sigma}\} \\ &= \mathbf{P}\{\sqrt{n}\frac{-0.005}{\sqrt{(0.505)(0.495)}} < \frac{X_n-\mu}{\sigma}\} \\ &= \mathbf{P}\{-0.01\sqrt{n} < \frac{X_n-\mu}{\sigma}\} \\ &= 1 - \Phi(-0.01\sqrt{n}) = \Phi(0.01\sqrt{n})\end{aligned}$$

Since $\Phi(1.645) = 0.95$, and Φ is an increasing function,

$$\begin{aligned}1.645 &< 0.01\sqrt{n} \qquad \text{or} \\ n &> 27,060\end{aligned}$$

Example – continued

☞ Suppose that Acme intends to survey 1000 people. What is the probability that over half voice the opinion to keep Hail to the Victors?

☞ From the previous, we have $n = 1000$, so the probability that over half the people surveyed want to keep the song is

$$\Phi(0.01\sqrt{1000}) \approx \Phi(0.32) \approx 0.6255$$

Polling

Example

Rasmussen Reports conducted their final Presidential Tracking poll for Election 2008, reporting on November 3 that Barack Obama held a 52% to 46% lead over John McCain. They noted (in the fine print)

- The margin of sampling error – for the full sample of 3000 likely voters – is ±2% with a 95% level of confidence.

What does this mean?

Polling

☞ The Rasmussen poll is trying to determine the actual percentage of voters who would vote for Obama/McCain in the day before the election. Let p be the actual of voters who really would vote for Obama on this day. Let $\overline{R}$ be percentage of polled voters who said they would vote for Obama.

☞ You are not likely to get the exact value of p in a small sample average $\overline{R}$. (This was the moral of the previous example.)

☞ Instead we replace an exact value $\overline{R}$ with a confidence interval $\overline{R} \pm \epsilon$ which we believe with high (i.e. 95%) likelihood contains the value of the unknown p. If ϵ is sufficiently small, this can be good enough.

Standardizing the poll outcome

☞ Let R_n be the binomial random variable giving the number of respondents who would vote for Obama in a randomly chosen sample of n. The actual (unknown) probability of voters would would vote for Obama is p. So,

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

☞ If the poll really provides an unbiased estimate, then for large enough n, R_n is approximated by the standard normal distribution.

Standardize R_n in order to approximate with Φ:

$$\frac{R_n - \mu}{\sigma} = \frac{R_n - np}{\sqrt{np(1-p)}}$$

Confidence Interval

☞ 95% of all outcomes lie within approximately two standard deviation, $\pm 2\sigma = 2$. (A closer estimate is ± 1.96.)

☞ We want n so that

$$\mathbf{P}\left\{-2 < \frac{R_n - np}{\sqrt{np(1-p)}} < 2\right\} \approx 0.95.$$

A little algebra

$$\mathbf{P}\left\{R_n - 2\sqrt{np(1-p)} < np < R_n + 2\sqrt{np(1-p)}\right\} \approx 0.95.$$

Divide through by n

$$\mathbf{P}\left\{\frac{R_n}{n} - 2\sqrt{\frac{p(1-p)}{n}} < p < \frac{R_n}{n} + 2\sqrt{\frac{p(1-p)}{n}}\right\} \approx 0.95.$$

Approximating the Confidence interval

$$\mathbf{P}\left\{\frac{R_n}{n} - 2\sqrt{\frac{p(1-p)}{n}} < p < \frac{R_n}{n} + 2\sqrt{\frac{p(1-p)}{n}}\right\} \approx 0.95.$$

☞ We want to bound

$$\left|\frac{R_n}{n} - p\right| \le 2\sqrt{\frac{p(1-p)}{n}}$$

We don't know p, but $p(1-p) = p - p^2$ has a maximum when $p = 0.5$ and value of 0.25. So we replace $p(1-p) = 0.25$:

$$\left|\frac{R_n}{n} - p\right| \le 2(0.5)\sqrt{\frac{1}{n}} = \sqrt{\frac{1}{n}}.$$

So,

$$\mathbf{P}\left\{\frac{R_n}{n} - \sqrt{\frac{1}{n}} < p < \frac{R_n}{n} + \sqrt{\frac{1}{n}}\right\} \approx 0.95.$$

Confidence Interval

$$\mathbf{P}\left\{\frac{R_n}{n} - \sqrt{\frac{1}{n}} < p < \frac{R_n}{n} + \sqrt{\frac{1}{n}}\right\} \approx 0.95.$$

☞ This shows that by increasing our sample size n, we can decrease the interval of 95% confidence, regardless of the actual probability p.

Sample Size	Confidence Interval (%)
10	± 0.32
100	± 0.1
1000	± 0.32
3000	± 0.018
10,000	± 0.01

☞ If the Rasmussen Poll was really unbiased, then we can say with 95% confidence that on November 3, 2008, 52% $\pm$ 2% of the voting population favored Obama.

The next day Obama gathered 52% of the votes.