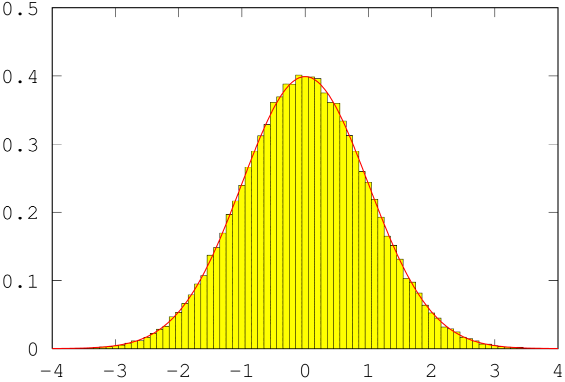
Continuous Expectation and Variance, the Law of Large Numbers, and the Central Limit Theorem
18.05 Spring 2014



Expected value

Expected value: measure of location, central tendency

X continuous with range $[a, b]$ and pdf $f(x)$:

$$E(X) = \int_a^b x f(x)\, dx.$$

X discrete with values $x_1, \ldots, x_n$ and pmf $p(x_i)$:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i).$$

View these as essentially the same formulas.

Variance and standard deviation

Standard deviation: measure of spread, scale

For *any* random variable X with mean μ

$$\mathrm{Var}(X) = E((X-\mu)^2), \qquad \sigma = \sqrt{\mathrm{Var}(X)}$$

X continuous with range $[a, b]$ and pdf $f(x)$:

$$\mathrm{Var}(X) = \int_a^b (x-\mu)^2 f(x)\, dx.$$

X discrete with values $x_1, \ldots, x_n$ and pmf $p(x_i)$:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} (x_i-\mu)^2 p(x_i).$$

View these as essentially the same formulas.

Properties

Properties: (the same for discrete and continuous)

1. $E(X + Y) = E(X) + E(Y)$.

2. $E(aX + b) = aE(X) + b$.

3. If X and Y are independent then
$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

4. $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

5. $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

Board question

The random variable X has range [0,1] and pdf cx^2.

(a) Find c.

(b) Find the mean, variance and standard deviation of X.

(c) Find the median value of X.

(d) Suppose $X_1, \ldots X_{16}$ are independent identically-distributed copies of X. Let $\overline{X}$ be their average. What is the standard deviation of $\overline{X}$?

(e) Suppose $Y = X^4$. Find the pdf of Y.

answer: *See next slides.*

Solution

(a) Total probability is 1: $\displaystyle\int_0^1 cx^2\,dx = 1 \;\Rightarrow\; \boxed{c = 3}$.

(b) $\mu = \int_0^1 3x^3\,dx = 3/4$.
$\sigma^2 = (\int_0^1 (x-3/4)^2\,3x^2\,dx) = \frac{3}{5} - \frac{9}{8} + \frac{9}{16} = \frac{3}{80}$.
$\sigma = \sqrt{3/80} = \frac{1}{4}\sqrt{3/5} \approx .194$

(c) Set $F(q_{0.5}) = 0.5$, solve for $q_{0.5}$: $F(x) = \displaystyle\int_0^x 3u^2\,du = x^3$. Therefore, $F(q_{0.5}) = q_{0.5}^3 = .5$. We get, $\boxed{q_{0.5} = (0.5)^{1/3}}$.

(d) Because they are independent
$\text{Var}(X_1 + \ldots + X_{16}) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_{16}) = 16\text{Var}(X)$.
Thus, $\text{Var}(\overline{X}) = \frac{16\text{Var}(X)}{16^2} = \frac{\text{Var}(X)}{16}$. Finally, $\sigma_{\overline{X}} = \boxed{\dfrac{\sigma_X}{4} = 0.194/4}$.

Solution continued

(e) Method 1 use the cdf:
$F_Y(y) = P(X^4 < y) = P(X < y^{\frac{1}{4}}) = F_X(y^{\frac{1}{4}}) = y^{\frac{3}{4}}$.
Now differentiate. $f_Y(y) = F_Y'(y) = \boxed{\frac{3}{4}y^{-\frac{1}{4}}}$.

Method 2 use the pdf: We have

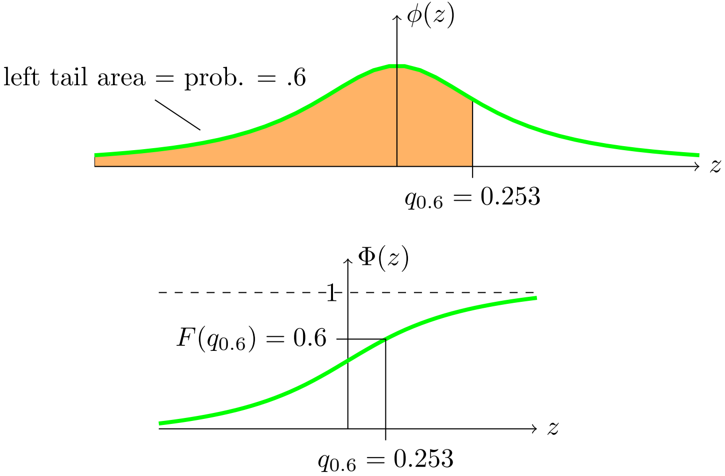
$$y = x^4 \;\Rightarrow\; dy = 4x^3\,dx \;\Rightarrow\; \frac{dy}{4y^{3/4}} = dx$$

This implies $$f_X(x)\,dx = f_X(y^{1/4})\frac{dy}{4y^{3/4}} = \frac{3y^{2/4}\,dy}{4y^{3/4}} = \frac{3}{4y^{1/4}}\,dy$$

Therefore $$f_Y(y) = \frac{3}{4y^{1/4}}$$

Quantiles

Quantiles give a measure of **location.**

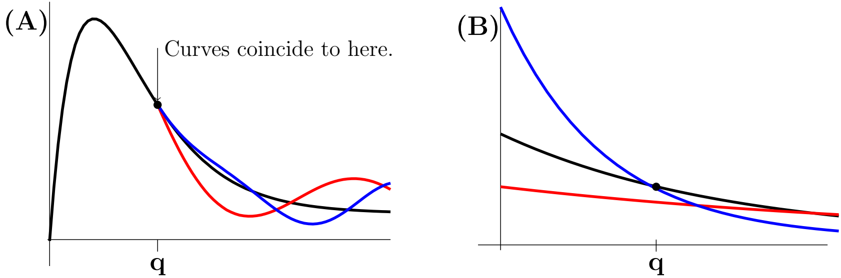


$q_{0.6}$: left tail area $= 0.6 \;\Leftrightarrow\; F(q_{0.6}) = 0.6$

Concept question

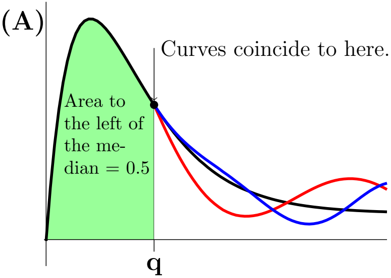
Each of the curves is the density for a given random variable. The median of the black plot is always at q. Which density has the greatest median?

1. Black
2. Red
3. Blue
4. All the same
5. Impossible to tell



answer: *See next frame.*

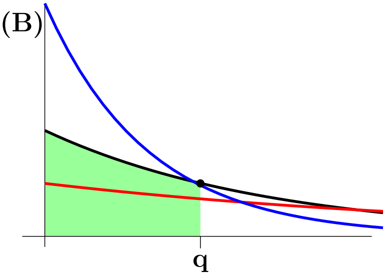
Solution



Plot A: $\boxed{\text{4. All three medians are the same.}}$ Remember that probability is computed as the area under the curve. By definition the median q is the point where the shaded area in Plot A .5. Since all three curves coincide up to q. That is, the shaded area in the figure is represents a probability of .5 for all three densities.

Continued on next slide.

Solution continued



Plot B: $\boxed{\text{2. The red density has the greatest median.}}$ Since q is the median for the black density, the shaded area in Plot B is .5. Therefore the area under the blue curve (up to q) is greater than .5 and that under the red curve is less than .5. This means the median of the blue density is to the left of q (you need less area) and the median of the red density is to the right of q (you need more area).

Law of Large Numbers (LoLN)

- Informally: An average of many measurements is more accurate than a single measurement.

- Formally: Let X_1, X_2, ... be i.i.d. random variables all with mean μ and standard deviation σ.
  Let
  $$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
  Then for any (small number) a, we have
  $$\lim_{n\to\infty} P(|\overline{X}_n - \mu| < a) = 1.$$

- **No guarantees but:** By choosing n large enough we can make $\overline{X}_n$ as close as we want to μ with probability close to 1.

Concept Question: Desperation

- You have \$100. You need \$1000 by tomorrow morning.
- Your only way to get it is to gamble.
- If you bet \$k, you either win \$k with probability p or lose \$k with probability $1 - p$.

Maximal strategy: Bet as much as you can, up to what you need, each time.
Minimal strategy: Make a small bet, say \$5, each time.

1. If $p = 0.45$, which is the better strategy?
   (a) Maximal (b) Minimal (c) They are the same

2. If $p = 0.8$, which is the better strategy?
   (a) Maximal (b) Minimal (c) They are the same

answer: *On next slide*

Solution to previous two problems

answer: If $p = 0.45$ use maximal strategy; If $p = 0.8$ use minimal strategy. If you use the minimal strategy the law of large numbers says your average winnings per bet will almost certainly be the expected winnings of one bet. The two tables represent $p = 0.45$ and $p = 0.8$ respectively.

Win	-10	10
p	0.55	0.45

Win	-10	10
p	0.2	0.8

The expected value of a \$5 bet when $p = 0.45$ is -\$0.50 Since on average you will lose \$0.50 per bet you want to avoid making a lot of bets. You go for broke and hope to win big a few times in a row. It's not very likely, but the maximal strategy is your best bet.

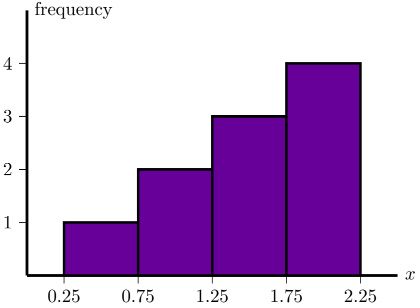
The expected value when $p = 0.8$ is \$3. Since this is positive you'd like to make a lot of bets and let the law of large numbers (practically) guarantee you will win an average of \$6 per bet. So you use the minimal strategy.

Histograms

Made by 'binning' data.

Frequency: height of bar over bin = number of data points in bin.

Density: area of bar is the fraction of all data points that lie in the bin. So, total area is 1.



Check that the total area of the histogram on the right is 1.

Board question

1. Make both a frequency and density histogram from the data below.

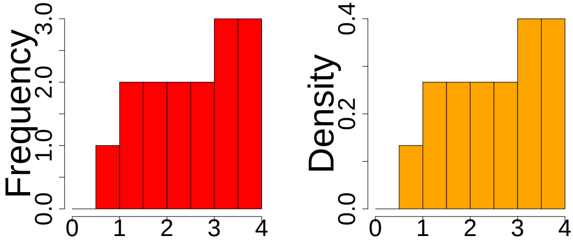
Use bins of width 0.5 starting at 0. The bins should be right closed.

1	1.2	1.3	1.6	1.6
2.1	2.2	2.6	2.7	3.1
3.2	3.4	3.8	3.9	3.9

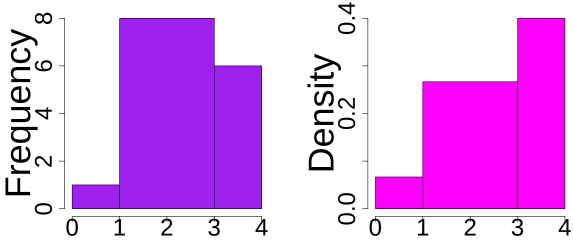
2. Same question using unequal width bins with edges 0, 1, 3, 4.

3. For question 2, why does the density histogram give a more reasonable representation of the data.

Solution



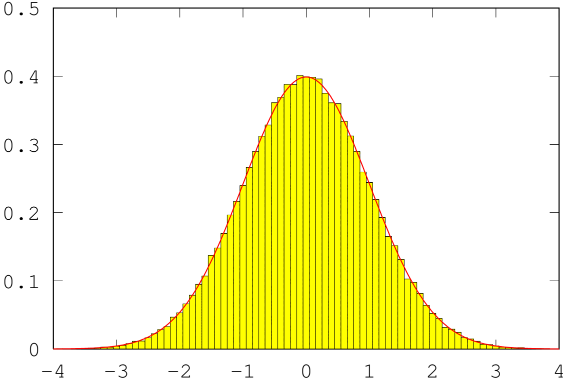
Histograms with equal width bins



Histograms with unequal width bins

LoLN and histograms

LoLN implies density histogram converges to pdf:



Histogram with bin width 0.1 showing 100000 draws from a standard normal distribution. Standard normal pdf is overlaid in red.

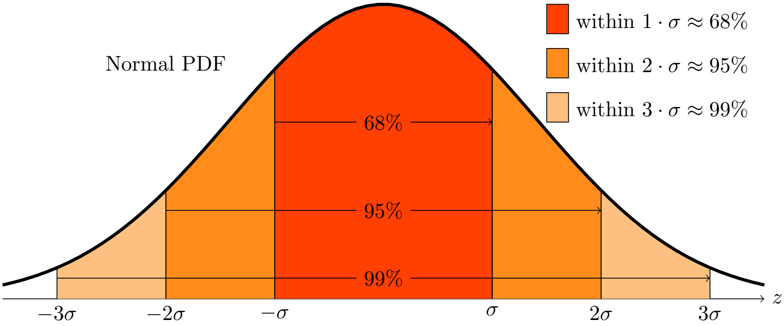
Standardization

Random variable X with mean μ and 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 use reserve Z to mean a standard normal random variable.

Concept Question: Standard Normal



1. $P(-1 < Z < 1)$ is
(a) 0.025 (b) 0.16 (c) 0.68 (d) 0.84 (e) 0.95

2. $P(Z > 2)$
(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:

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

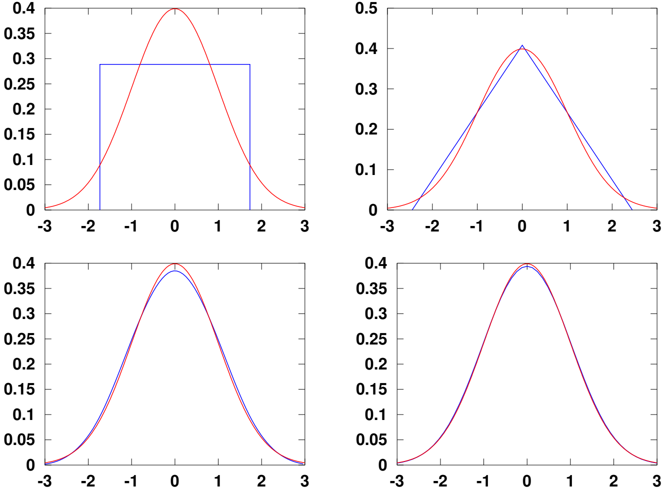
$$S_n \approx \mathrm{N}\left(n\mu, n\sigma^2\right)$$

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

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

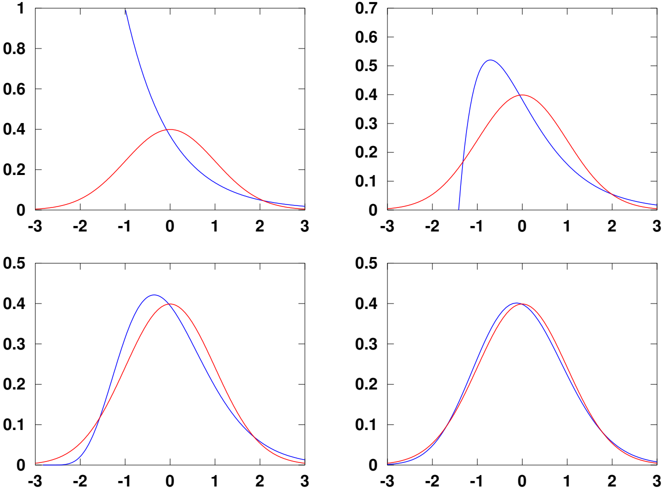
CLT: pictures

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



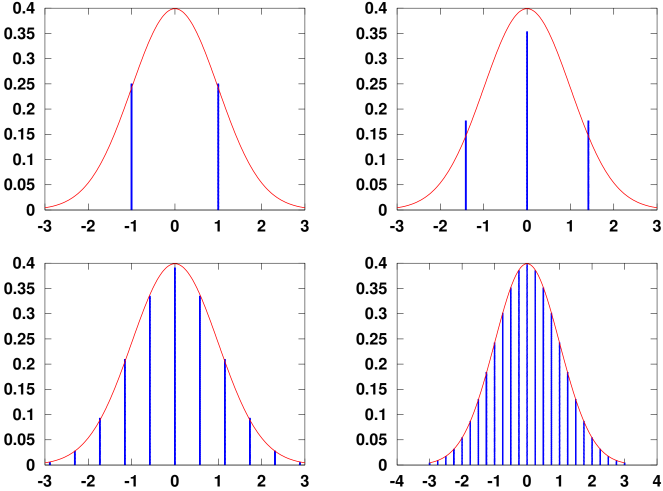
CLT: pictures 2

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



CLT: pictures 3

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



CLT: pictures 4

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

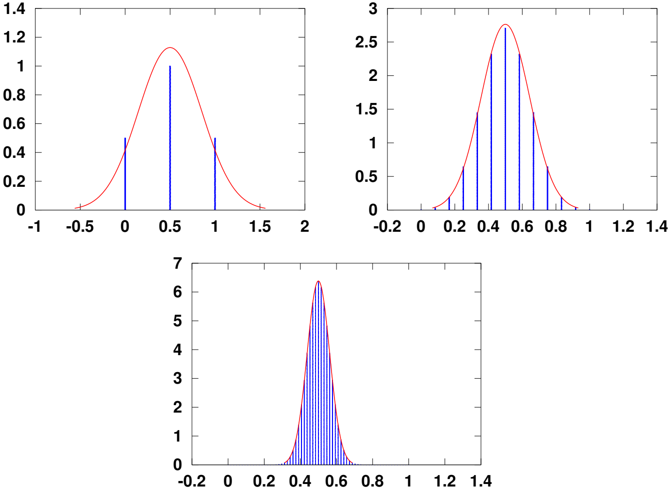


Table Question: Sampling from the standard normal distribution

As a table, produce a single random sample from (an approximate) standard normal distribution.

The table is allowed nine rolls of the 10-sided die.

Note: $\mu = 5.5$ and $\sigma^2 = 8.25$ for a single 10-sided die.

Hint: CLT is about averages.

answer: 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?

<u>**answer:**</u> 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
$\sigma^2_{\mathcal{A}} = 0.25/400$ 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

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

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.

<u>**answer:**</u> 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 $P(S < -5 \text{ or } S > 5) \approx P(Z < -1 \text{ or } Z > 1) \approx \boxed{0.32}$.