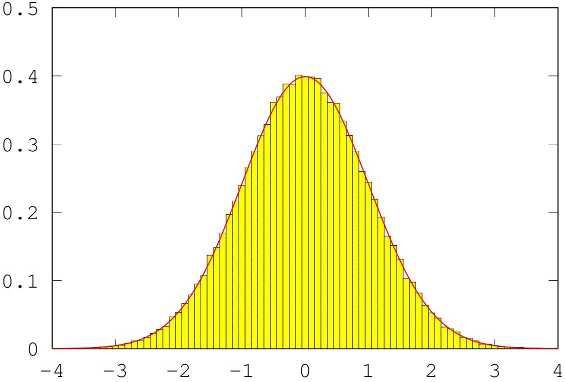
Continuous Expectation and Variance, the Law of Large Numbers, and the Central Limit Theorem
18.05 Spring 2014



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
   $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
4. $\text{Var}(aX + b) = a^2\text{Var}(X)$.
5. $\text{Var}(X) = E(X^2) - E(X)^2$.

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.

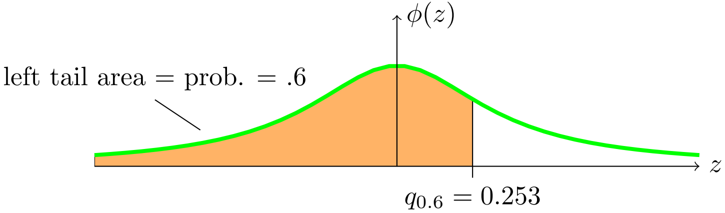
(c) Find the median value of X.

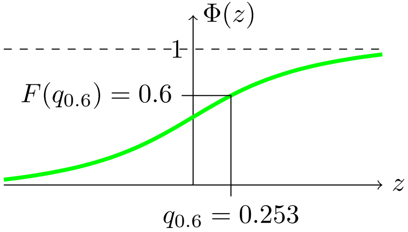
(d) Suppose $X_1, \ldots X_{16}$ are independent identically-distributed copies of X. Let $\overline{X}$ be their average. What is the standard deviation of $\overline{X}$?

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

Quantiles

Quantiles give a measure of **location.**



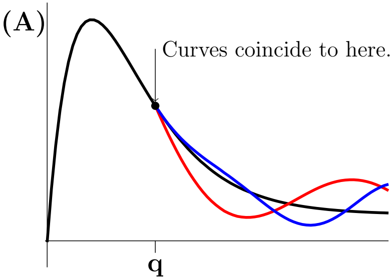


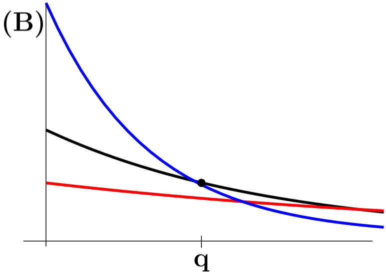
$q_{0.6}$: left tail area $= 0.6 \;\Leftrightarrow\; F(q_{0.6}) = 0.6$

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





Law of Large Numbers (LoLN)

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

- Formally: Let $X_1, X_2, \ldots$ 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

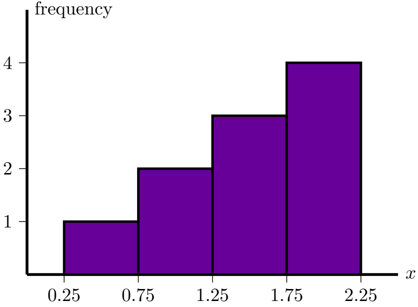
2. If $p = 0.8$, which is the better strategy?
   (a) Maximal (b) Minimal (c) They are the same

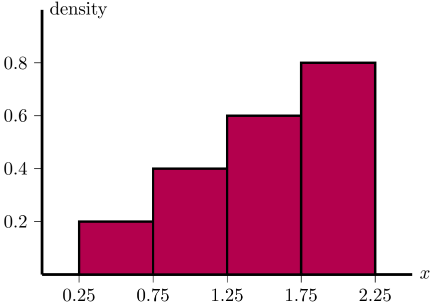
Histograms

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

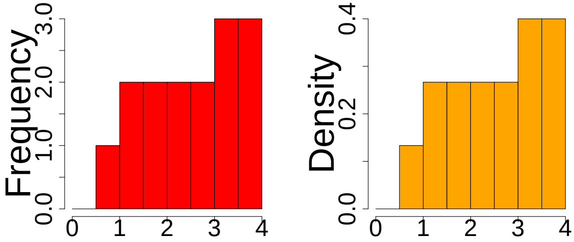
1. Make both a frequency and density histogram from the data below. Use bins of width 0.5 starting at 0. The bins should be right closed.

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

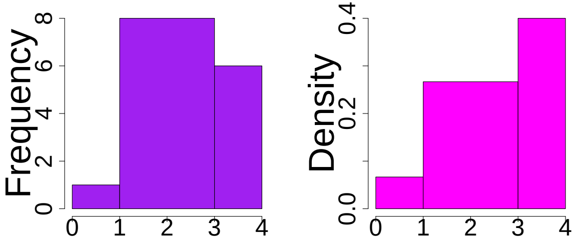
2. Same question using unequal width bins with edges 0, 1, 3, 4.

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

Solution



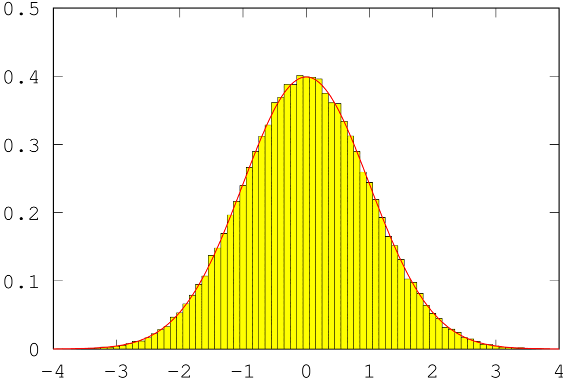
Histograms with equal width bins



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.

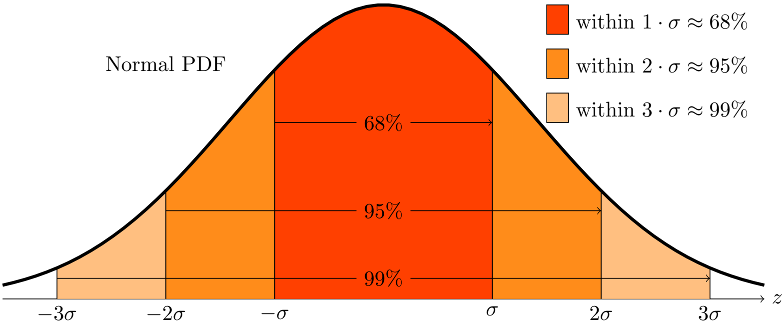
Standardization

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

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:

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

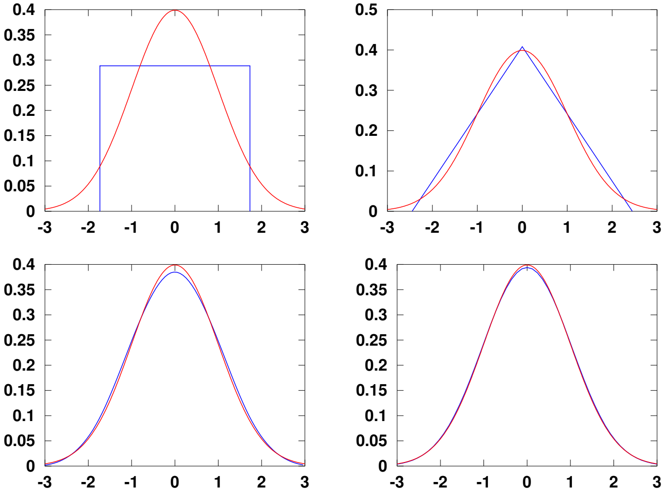
$$S_n \approx \mathrm{N}\left(n\mu, n\sigma^2\right)$$

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

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

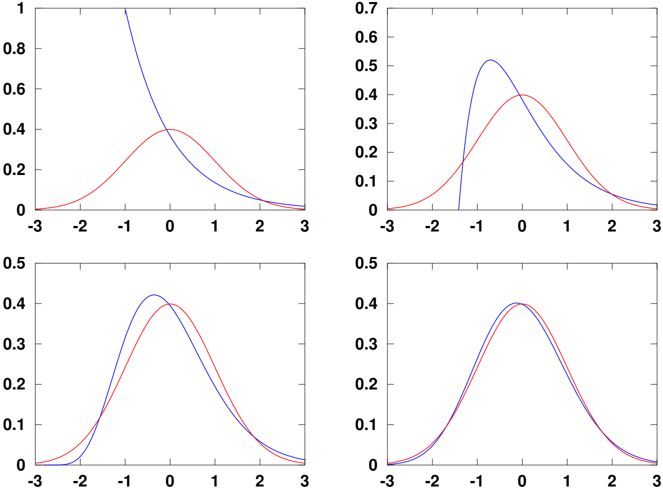
CLT: pictures

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



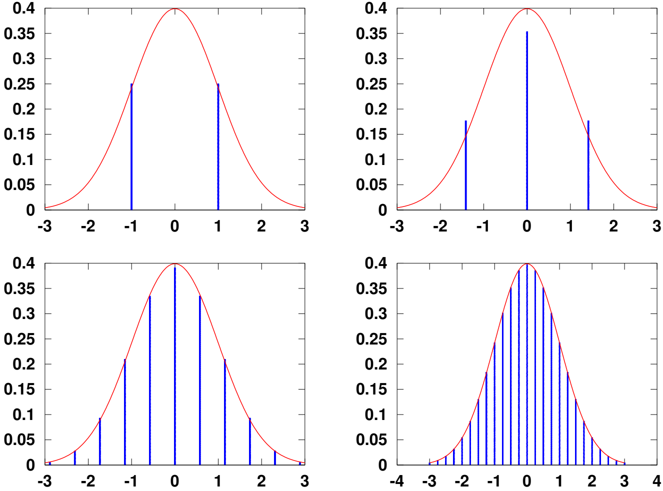
CLT: pictures 2

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



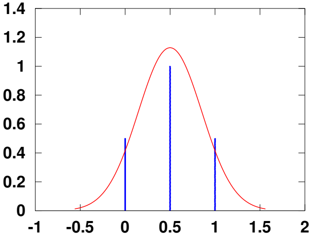
CLT: pictures 3

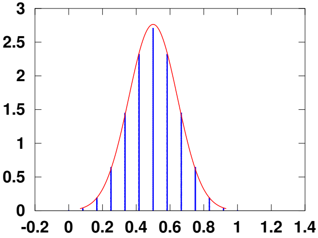
The standardized average of n i.i.d. Bernoulli(0.5) random variables with $n = 1, 2, 12, 64$.



CLT: pictures 4

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





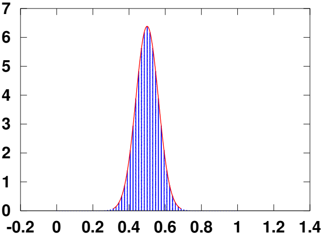


Table Question: Sampling from the standard normal distribution

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.

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?

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.

18.05 Introduction to Probability and Statistics
Spring 2014