Applied Biostatistics for Life Sciences

Dmitri D. Pervouchine

Skolkovo Institute for Science and Technology

Part 1: Exploratory Analyses and Probability

Contents

Sample vs. Population

- The entire group of objects about which information is wanted is called the **population**
- Individual members are called **units**
- A small part of the population is actually available for observation; it is called **sample**
- A **statistic** is a function of the sample

Problem 1.1

Fill in the missing words to the quote: "Statistical methods may be described as methods for drawing conclusions about . . . . . based on . . . . . computed from the . . . . . .

(A) *statistics, samples, populations*

(B) *populations, parameters, samples*

(C) *statistics, parameters, samples*

(D) *parameters, statistics, populations*

(E) *populations, statistics, samples*

Solution

(E) populations, statistics, samples

The types of variables

- **Numeric** variables have values that describe a measurable quantity as a number
  - A **continuous** variable can take any value between a certain set of real numbers
  - A **discrete** variable can take a value from a set of distinct (usually, integer) values
- **Categorical** variables have values that belong to categories
  - An **ordinal** variable has values that can be logically ordered or ranked
  - Values of a **nominal** cannot be logically ordered

Measures of center

- $\{X_1, X_2, \ldots, X_n\}$ is the sample
- Assume the sample is sorted, i.e., $X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_n$
- Statistic $= f(X_1, X_2, \ldots, X_n)$
- Measures of center
  - **Median** is the $50\%^{\text{th}}$ observation; $\text{MED} = X_{[\frac{n+1}{2}]}$
  - Note that if $\frac{n+1}{2}$ is not an integer, the definition of the median is a matter of convention
  - **Mean** is the center of mass; $\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$
  - **Mode** is the most frequent observation

Percentiles

- A **percentile** is a measure indicating the value below which a given percentage of observations in a group of observations fall
- $X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_n$ is the sorted sample
- **Upper quartile** is the 75th percentile; $\mathrm{UQ} = X_{[0.75(n+1)]}$
- **Lower quartile** is the 25th percentile; $\mathrm{LQ} = X_{[0.25(n+1)]}$
- Median is the 50th percentile
- *Note that if* $[0.25(n+1)]$ *is not an integer, the definition of UQ and LQ is a matter of convention*

Measures of spread

- $X_1 \leq X_2 \leq X_3 \leq \ldots \leq X_n$ is the sample
- **Interquartile range** $IQR = UQ - LQ$
- RANGE $= X_{max} - X_{min}$
- **Variance** $= s_X^2 = \frac{(X_1-\overline{X})^2+(X_2-\overline{X})^2+\cdots+(X_n-\overline{X})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$
- Variance is the moment of inertia
- **Standard deviation** $s_X = \sqrt{Variance}$

Outliers

- An **outlier** is an observation point that is distant from other observations
- An outlier is *defined to be* the value in the sample that differs from the nearest quartile by more than $1.5IQR$
- Susceptible to outliers: mean, variance, standard deviation, range
- Not susceptible to outliers: quartiles, median, interquartile range

Problem 2.1

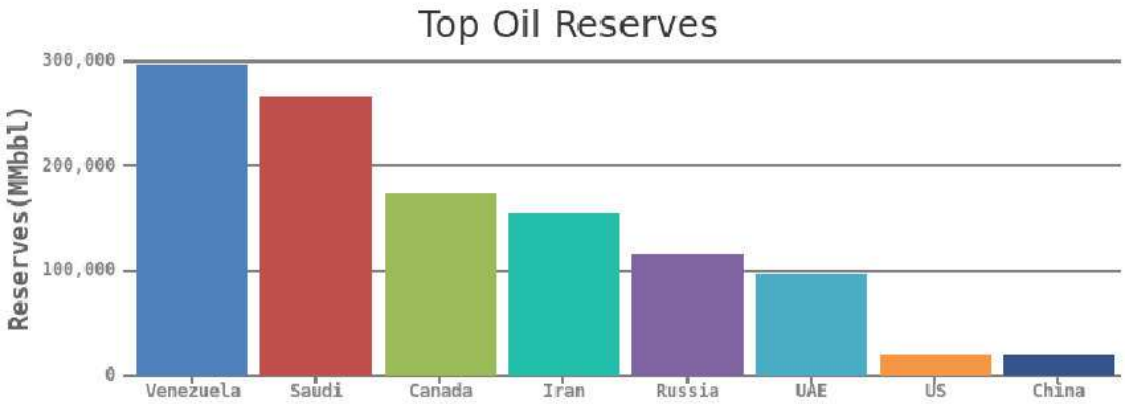
Find mean, median, mode, interquartile range, and standard deviation for the following sample: $-2, 1, 1, 2, 3, 4, 4, 4, 5, 6$

Solution

- $\overline{X} = \frac{-2+1+1+2+3+4+4+4+5+6}{10} = 2.8$
- $s_X^2 = \frac{(-2-2.8)^2+(1-2.8)^2+\cdots+(6-2.8)^2}{9} = 5.51$
- $s_X = 2.35$
- $\frac{n+1}{4} = \frac{11}{4} =$ *between 2nd and 3rd obs*
- $\frac{3(n+1)}{4} = \frac{33}{4} =$ *between 8th and 9th obs*
- *The median is the average of* X_5 *and* X_6 *in the sorted sample*
- $MED = \frac{3+4}{2} = 3.5$
- *Mode=4*
- $LQ = (1+1)/2 = 1$
- $UQ = (4+5)/2 = 4.5$

Barplot

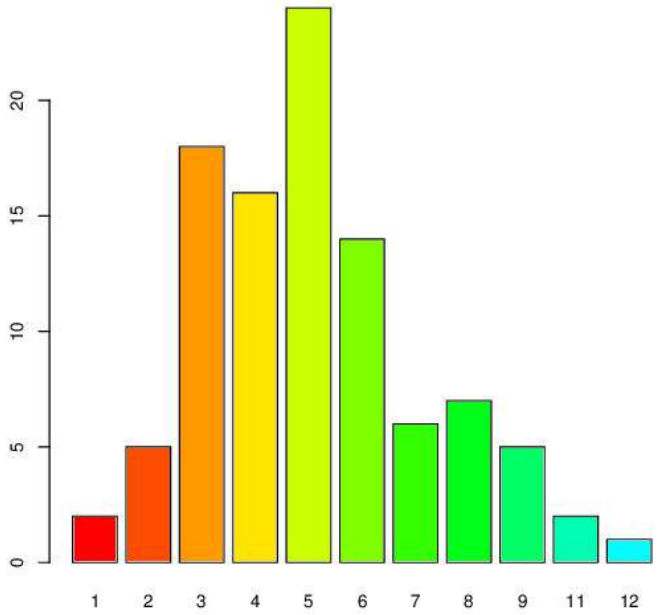
A **bar chart** or **bar graph** or **bar plot** is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent



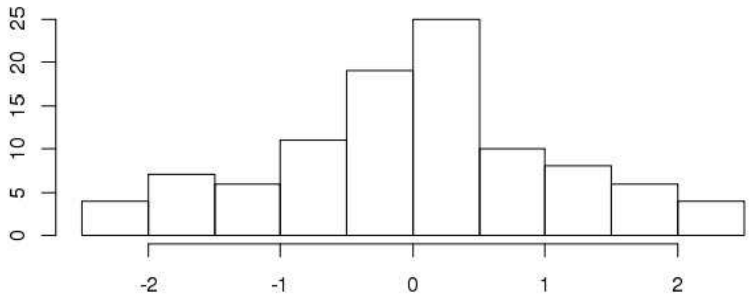
Barplot

Very often bar plot represents frequencies of observations

11 6 3 5 5 4 5 7 6 5 8 6 6 6 3 3 7 7 7 3 12 9 5 5 4
8 6 8 4 4 5 9 3 4 6 2 3 2 4 3 3 5 5 4 5 7 5 4 3 6
5 4 6 6 3 3 3 9 8 5 3 2 2 4 5 3 2 7 4 5 5 8 8 5 3
4 5 6 4 5 6 5 5 1 3 1 9 4 5 5 4 3 8 5 9 4 3 6 6 11



Histogram

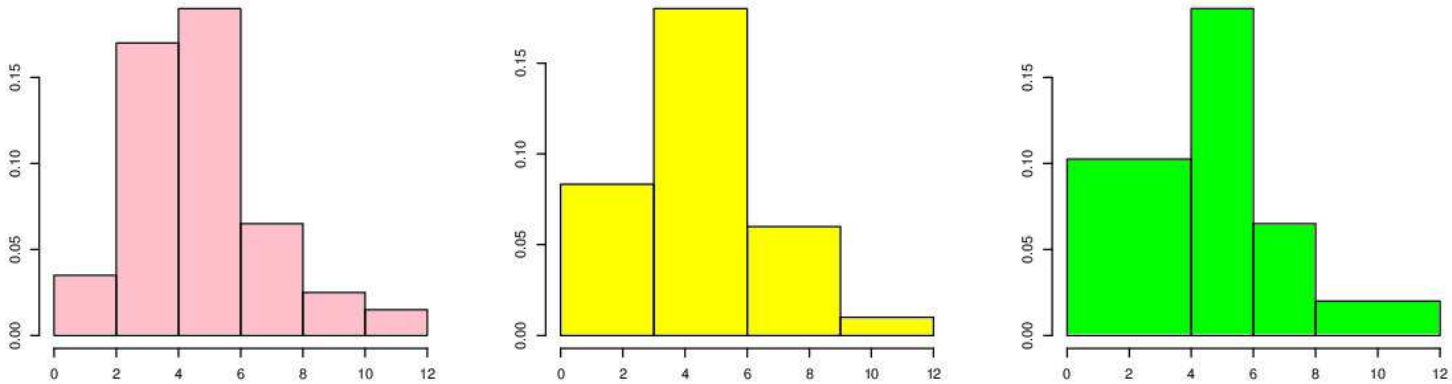


- The first step is to "bin" the range of values
- If the bins are of equal size, the *height* of the rectangle over the bin is proportional to the frequency
- If the bins are not of equal size, the *area* of the rectangle is proportional to the frequency
- The vertical axis is not frequency but density: the number of cases per unit of the variable on the horizontal axis

Histogram

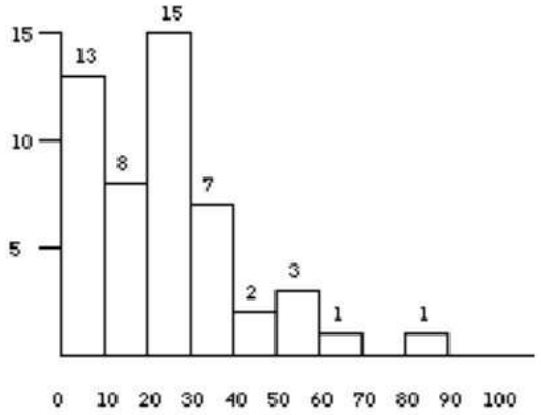
Histogram depends on the binning:

11 6 3 5 5 4 5 7 6 5 8 6 6 6 3 3 7 7 7 3 12 9 5 5 4
8 6 8 4 4 5 9 3 4 6 2 3 2 4 3 3 5 5 4 5 7 5 4 3 6
5 4 6 6 3 3 3 9 8 5 3 2 2 4 5 3 2 7 4 5 5 8 8 5 3
4 5 6 4 5 6 5 5 1 3 1 9 4 5 5 4 3 8 5 9 4 3 6 6 11



Problem 2.2

The following is a histogram showing the actual frequency of the closing prices on the New York exchange of a particular stock. Based on the above frequency histogram for New York Stock exchange, what is the class that contains the 80th percentile ?



(A) *20-30*

(B) *10-20*

(C) *40-50*

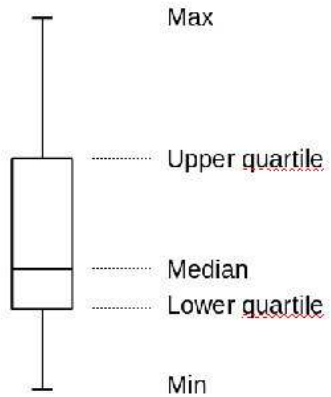
(D) *50-60*

(E) *30-40*

Solution

Correct: **(E)**

Boxplot



- A boxplot is a way of graphically depicting groups of numerical data through their quartiles
- Box plots also have lines extending the boxes (whiskers) indicating variability outside the upper and lower quartiles
- Outliers are plotted as individual points

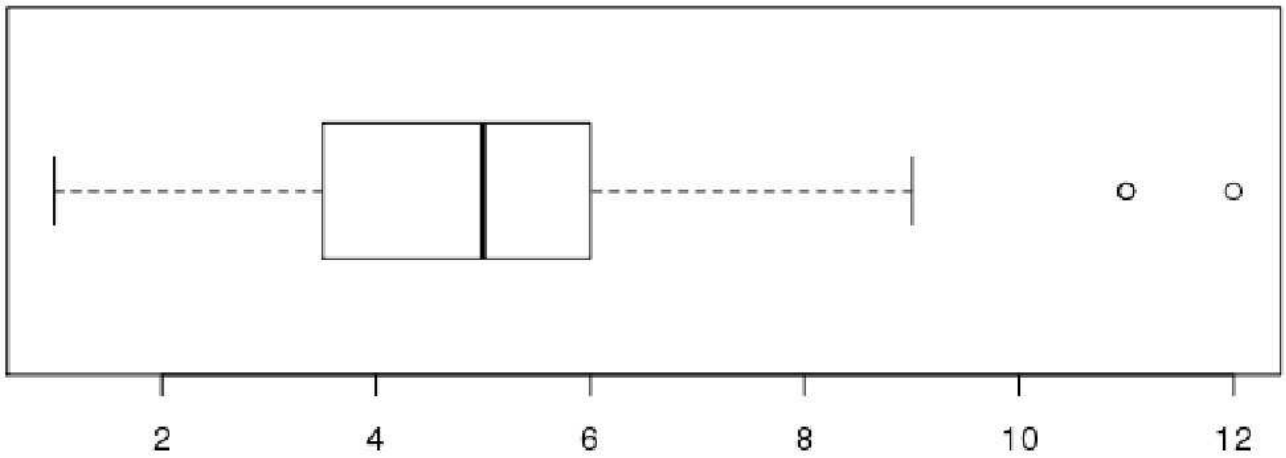
Problem 2.3

Draw the boxplot for the following sample

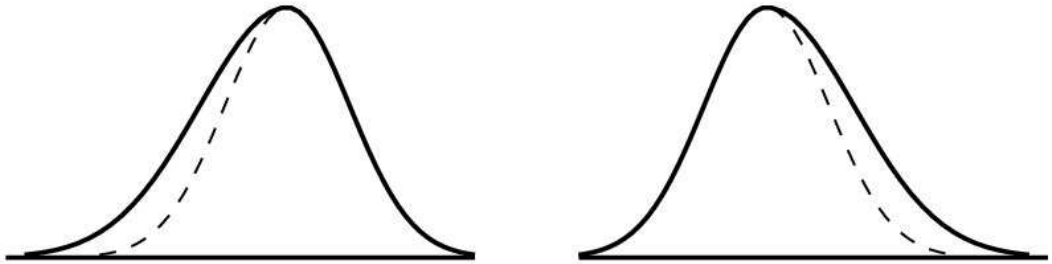
11 6 3 5 5 4 5 7 6 5 8 6 6 6 3 3 7 7 7 3 12 9 5 5 4
8 6 8 4 4 5 9 3 4 6 2 3 2 4 3 3 5 5 4 5 7 5 4 3 6
5 4 6 6 3 3 3 9 8 5 3 2 2 4 5 3 2 7 4 5 5 8 8 5 3
4 5 6 4 5 6 5 5 1 3 1 9 4 5 5 4 3 8 5 9 4 3 6 6 11

Solution

Min	*LQ*	*Median*	*UQ*	*Max*
1.00	*3.75*	*5.00*	*6.00*	*12.00*



Skewness



Skewed to the left | Skewed to the right

- skewed to the left: the left tail is longer
- skewed to the right: the right tail is longer
- Relationship between mean and median
  - skewed to the left: mean $<$ median
  - skewed to the right: mean $>$ median
- Outliers from the right: skewed to the right

Experiment and variables

$$y = f(x, \alpha_1, \alpha_2, \ldots, \alpha_n)$$

- x is **explanatory variable**; the experimenter sets its values
- y is **response variable**; it is the measured response
- $\alpha_2, \alpha_2, \ldots, \alpha_n$ are **confounding variables**
- **Confounding variables also affect the response variable, but their values cannot be controlled**

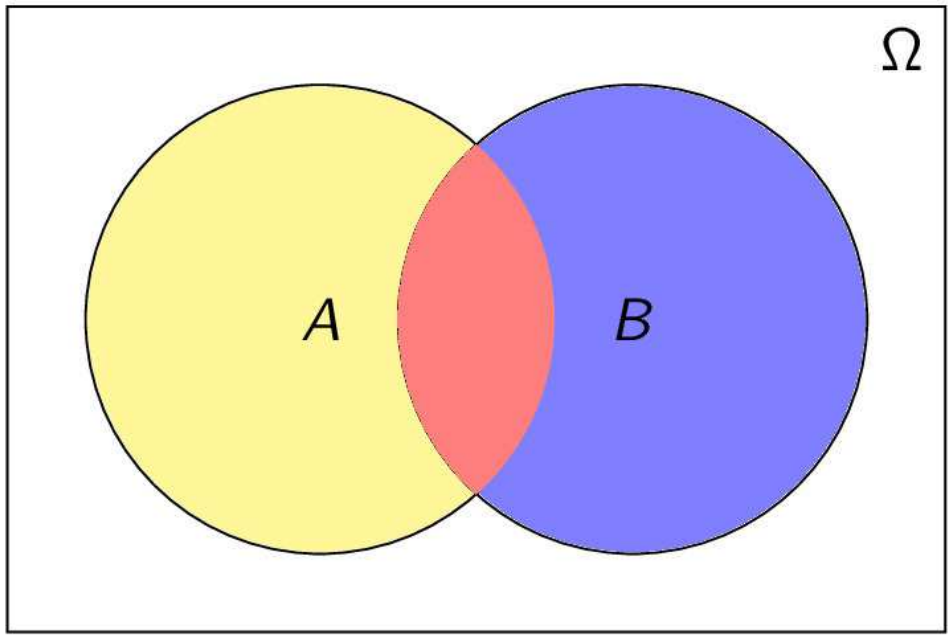
Control over confounding variables

- Blocking = divide units into blocks
  - within block elements are similar
  - between blocks elements are different
- Control group
  - Subtract baseline levels
  - measure the change instead of the absolute value
- Randomization
  - assign elements randomly into the treatment and control group
  - hope that the effects of confounding factors will "average out"

Probability

Probability is a *normalized measure*

- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\Omega) = 1$



Problem 3.1

Given that $\mathrm{P}(A) = 0.6$ *and* $\mathrm{P}(B) = 0.7$, *which of the following* **cannot** *be true? [Recall that* $\cup = OR$; $\cap = AND$*]*

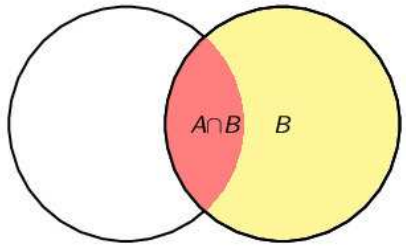
1. $\mathrm{P}(A \cap B) = 0.5$
2. $\mathrm{P}(A \cup B) = 0.9$
3. $\mathrm{P}(A \cap B) = 0.2$
4. $\mathrm{P}(A \cup B) = 0.4$
5. $\mathrm{P}(A \cap B) = 0.7$

Solution

Cannot be true:

3. $\mathrm{P}(A \cap B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cup B) \geq \mathrm{P}(A) + \mathrm{P}(B) - 1 = 0.3$
4. $\mathrm{P}(A \cup B) \geq \max\{\mathrm{P}(A), \mathrm{P}(B)\} = 0.7$
5. $\mathrm{P}(A \cap B) \leq \min\{\mathrm{P}(A), \mathrm{P}(B)\} = 0.6$

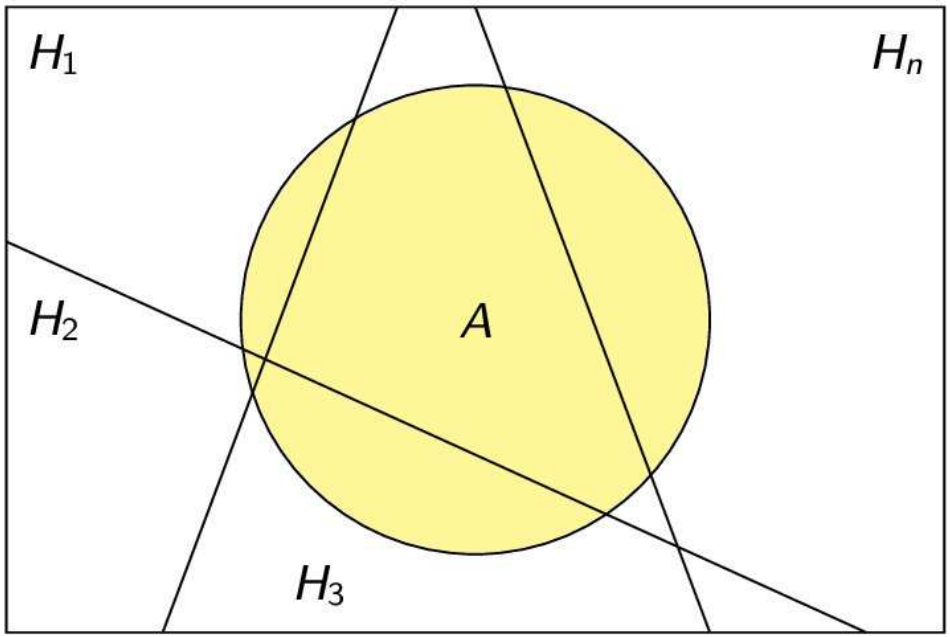
Conditional Probability



- A and B are called independent if $\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$
- $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$
- $\mathrm{P}(A|B) =$ the fraction of A in B
- A and B are independent iff $\mathrm{P}(A|B) = P(A)$

Law of Total Probability and Bayes Formula

- $\mathrm{P}(A) = \mathrm{P}(A|H_1)\,\mathrm{P}(H_1) + \mathrm{P}(A|H_2)\,\mathrm{P}(H_2) + \ldots + \mathrm{P}(A|H_n)\,\mathrm{P}(H_n)$
- In particular, $\mathrm{P}(A) = \mathrm{P}(A|B)\,\mathrm{P}(B) + \mathrm{P}(A|\overline{B})\,\mathrm{P}(\overline{B})$

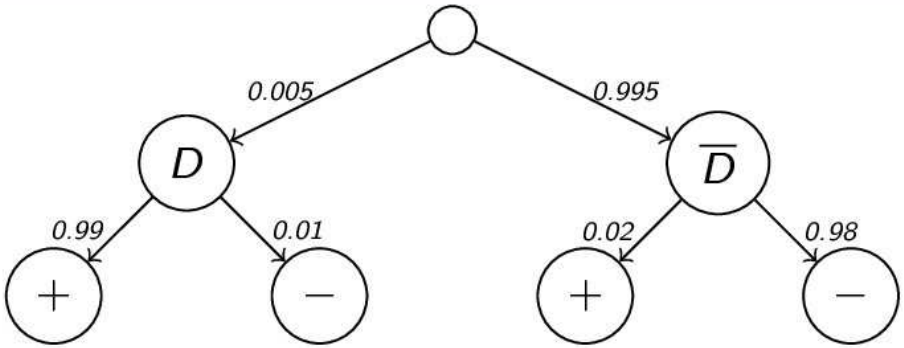


$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Problem 3.2

Suppose a certain drug test is 99% sensitive and 98% specific, that is, the test will correctly identify a drug user as testing positive 99% of the time and will correctly identify a non-user as testing negative 98% of the time. Let's assume a corporation decides to test its employees for opium use, and 0.5% of the employees use the drug. What is the probability that, given a positive drug test, an employee is actually a drug user?

Solution



$$P(D|+) = 0.005 \cdot 0.99/(0.005 \cdot 0.99 + 0.995 \cdot 0.02) = 0.1991$$

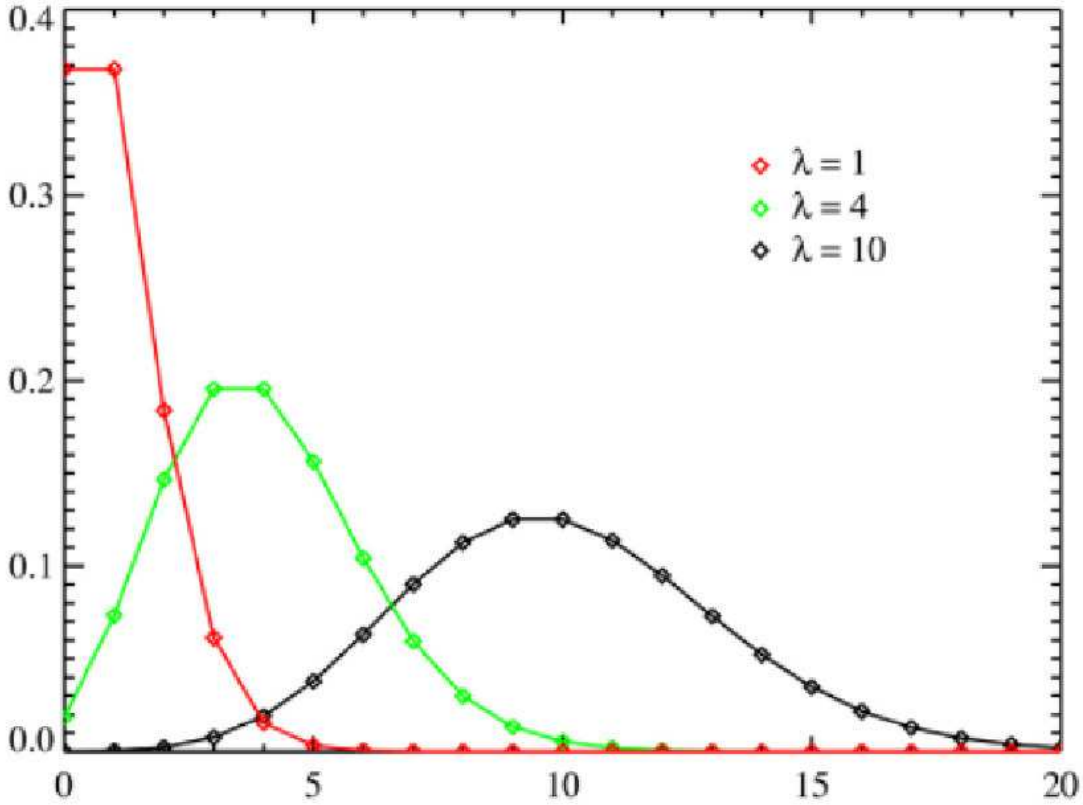
Distributions

- Discrete (Uniform, Binomial, Poisson, Geometric, Hypergeometric, Negative Binomial, ...)

- Continuous (Uniform, Normal, Exponential, Gamma, Chi-square, Student, Fisher, Dirichlet, ...)

Discrete Distributions

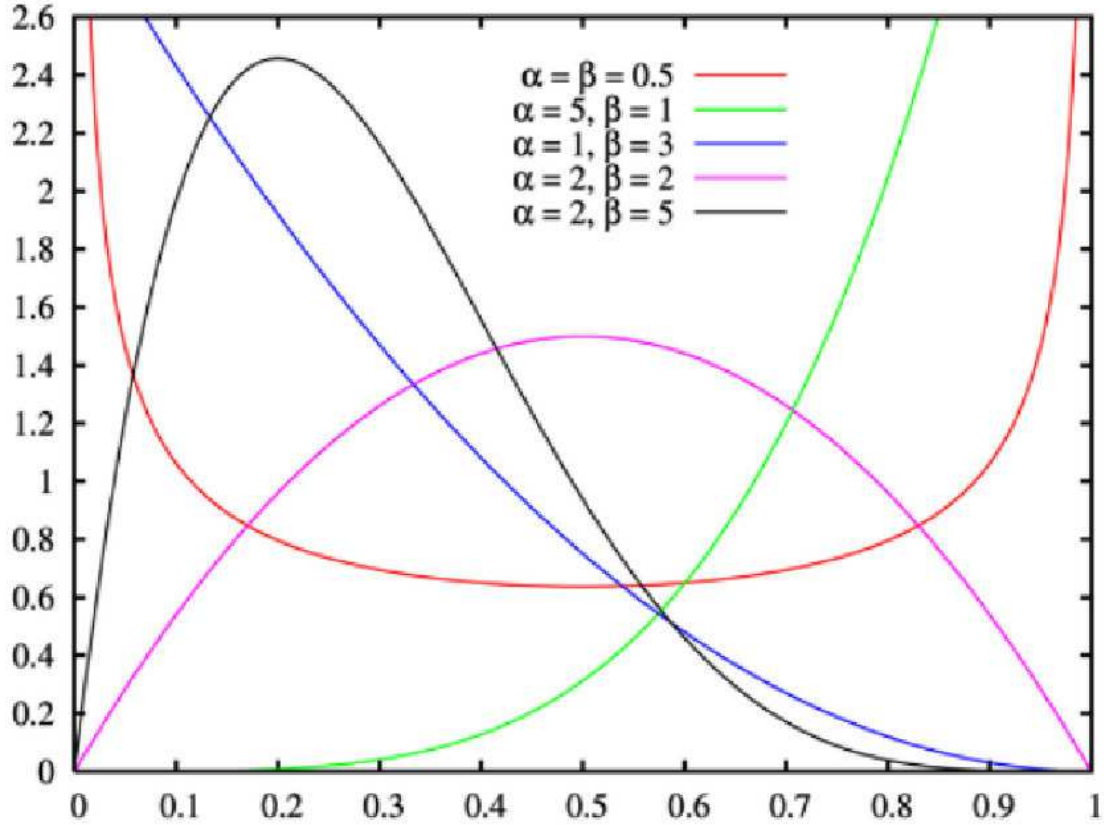
In a discrete distribution, probability is assigned to isolated values



Poisson distribution

Continuous Distributions

In a continuous distribution, probability density is smirred over a range of values



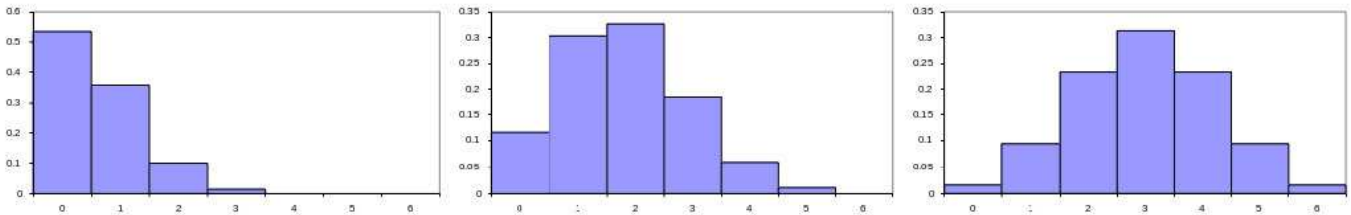
Beta distribution

Convention on the Notation for Random Numbers

- Random numbers will be denoted by capitals $X, Y, Z, \ldots$
- Regular (non-random numbers) will be denoted by small $x, y, z, \ldots$
- The expression "$X = 1$" makes no sense
- The expression $\mathrm{P}(X = 1) = 0.45$ does make sense

Binomial Distribution

- Binomial random number = the number of successes in n independent trials; p is the probability of success in one trial
- $\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, where $k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot k$



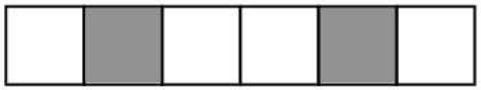
p=0.1 p=0.3 p=0.5

Problem 4.1

The probability that a certain machine will produce a defective item is 0.20. If a random sample of 6 items is taken from the output of this machine, what is the probability that there will be exactly 2 defectives in the sample?

Solution

$$P(1 \cap 2) = 0.2 \cdot 0.2 \cdot 0.8 \cdot 0.8 \cdot 0.8 \cdot 0.8 = 0.2^2 \cdot 0.8^4$$



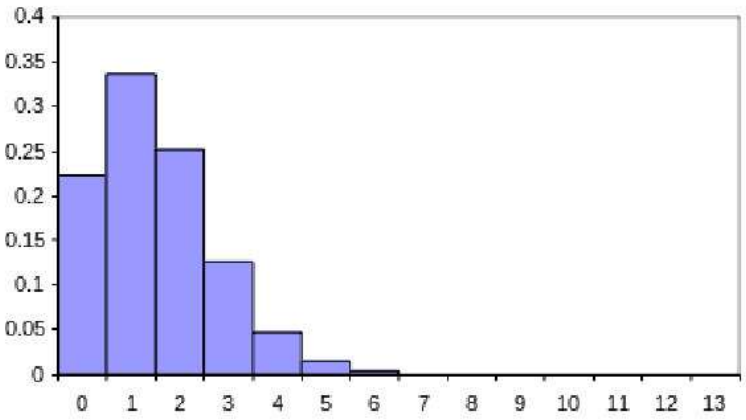
$$P(2 \cap 5) = 0.8 \cdot 0.2 \cdot 0.8 \cdot 0.8 \cdot 0.2 \cdot 0.8 = 0.2^2 \cdot 0.8^4$$

$$P = \binom{6}{2} 0.2^2 \cdot 0.8^4 = \frac{6 \cdot 5}{2 \cdot 1} 0.2^2 \cdot 0.8^4 \simeq 0.2458$$

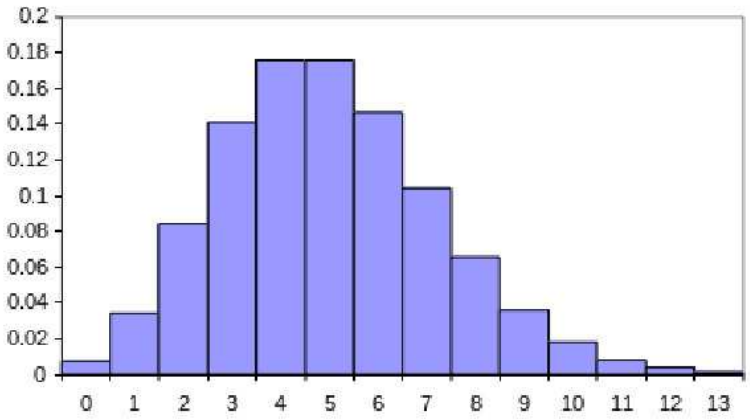
Poisson Distribution

Poisson random number = the number of rare events per unit of time or space; λ is the intensity parameter

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



λ=1.5



λ=5

Problem 4.2

The marketing manager of a company has noted that she usually receives 10 complaint calls during a week (consisting of five working days), and that the calls occur at random. Find the probability that she gets five such calls in one day.

Solution

- $X =$ *the number of complaint calls per day*
- $X \sim Poisson(\lambda = 2)$
- $\mathrm{P}(X = 5) = \frac{2^5}{5!} e^{-2} = 0.0361$

Problem 4.3

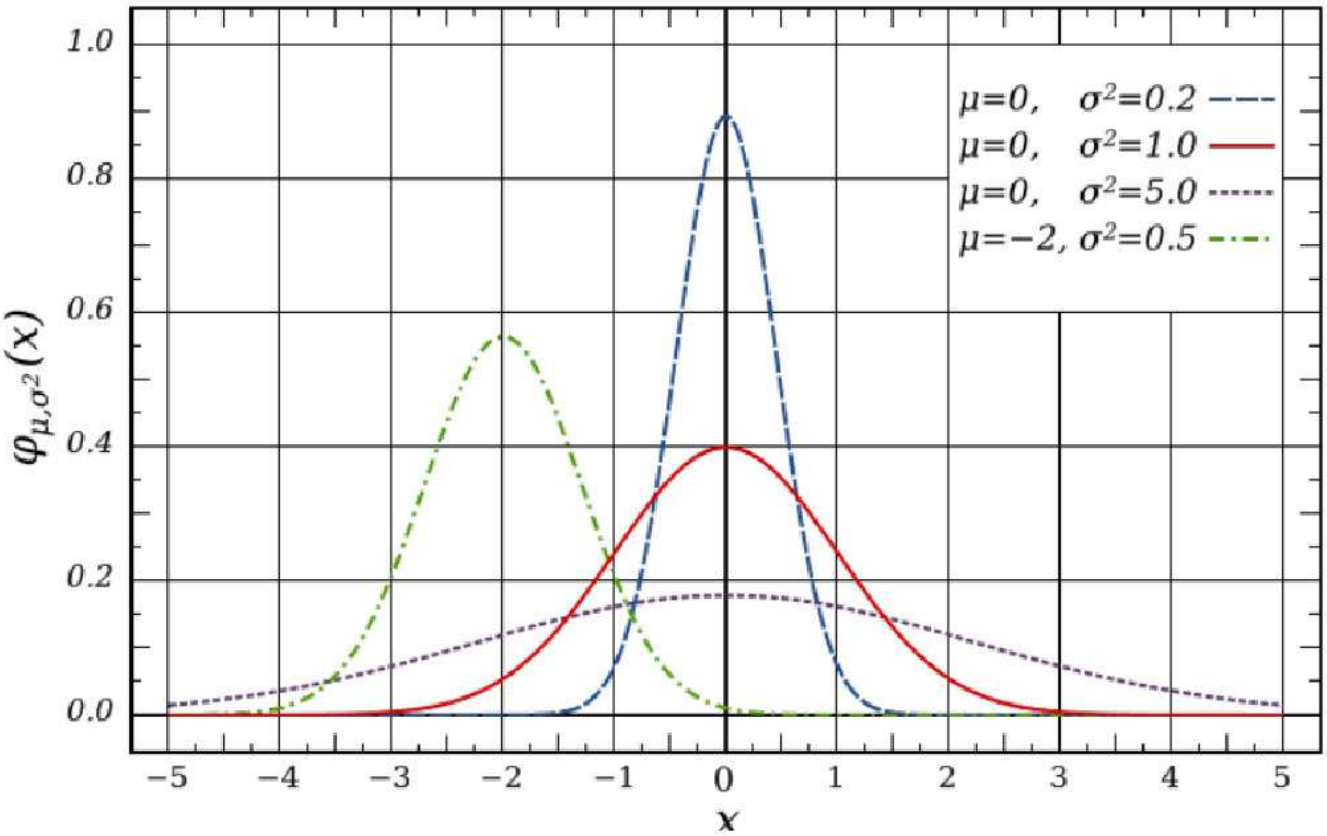
The rate at which a particular defect occurs in lengths of plastic film being produced by a stable manufacturing process is 4.2 defects per 75 meter length. A random sample of the film is selected and it was found that the length of the film in the sample was 25 meters. What is the probability that there will be at most 2 defects found in the sample?

Solution

- $X =$ *the number of defects per 25 meters*
- $X \sim Poisson(\lambda = 1.4)$
- $\mathrm{P}(X \leq 2) = e^{-1.4}\left(\frac{1.4^0}{0!} + \frac{1.4^1}{1!} + \frac{1.4^2}{2!}\right) = 0.8335$

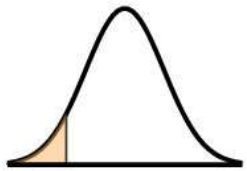
Normal Distribution

Normal random number = contribution of many independent random factors; μ = measure of center, σ = measure of spread.



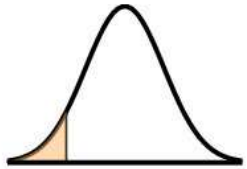
Cumulative Probability Tables

Standard normal distribution: $\mu = 0$ and $\sigma = 1$.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Cumulative Probability Tables



z	.00	.01	.02	.03	.04
-3.4	.0003	.0003	.0003	.0003	.0003
-3.3	.0005	.0005	.0005	.0004	.0004
-3.2	.0007	.0007	.0006	.0006	.0006
-3.1	.0010	.0009	.0009	.0009	.0008
-3.0	.0013	.0013	.0013	.0012	.0012
-2.9	.0019	.0018	.0018	.0017	.0016
-2.8	.0026	.0025	.0024	.0023	.0023
-2.7	.0035	.0034	.0033	.0032	.0031
-2.6	.0047	.0045	.0044	.0043	.0041
-2.5	.0062	.0060	.0059	.0057	.0055
-2.4	.0082	.0080	.0078	.0075	.0073
-2.3	.0107	.0104	.0102	.0099	.0096
-2.2	.0139	.0136	.0132	.0129	.0125
-2.1	.0179	.0174	.0170	.0166	.0162
-2.0	.0228	.0222	.0217	.0212	.0207

$$P(Z < -2.51) = 0.0060$$

Other Normal Distributions

- $Z \sim \mathcal{N}(0, 1)$
  - Mean $= 0$
  - Standard deviation $= 1$

- $X \sim \mathcal{N}(\mu, \sigma)$
- Note that some authors denote $\mathcal{N}(\mu, \sigma^2)$
  - Mean $= \mu$
  - Standard deviation $= \sigma$

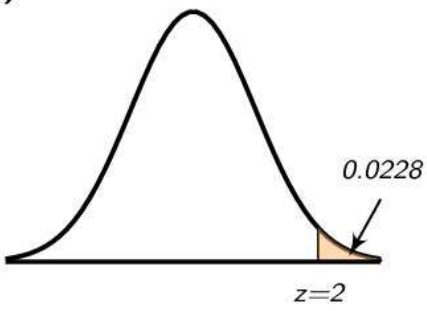
- $Z = (X - \mu)/\sigma$

Problem 4.4

The diameters of steel disks produced in a plant are normally distributed with a mean of 2.5 cm and standard deviation of 0.02 cm. What is the probability that a disk picked at random has a diameter greater than 2.54 cm?

Solution

- $X =$ *diameters of a randomly picked steel disk*
- $X \sim \mathcal{N}(\mu = 2.5, \sigma = 0.02)$
- $\mathrm{P}(X > 2.54) = \mathrm{P}(X - \mu > 2.54 - 2.5) = \mathrm{P}(\frac{X-\mu}{\sigma} > \frac{2.54-2.5}{0.02}) = \mathrm{P}(Z > 2)$



$$\mathrm{P}(Z > 2) = \mathrm{P}(Z < -2) = 0.0228$$

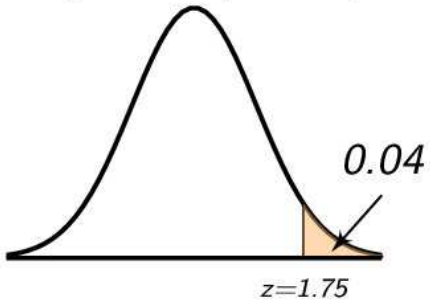
Problem 4.5

The height of an adult male is known to be normally distributed with a mean of 69 inches and a standard deviation of 2.5 inches. What is the height of the doorway such that 96 percent of the adult males can pass through it without having to bend?

Solution

- $X =$ *height of a randomly picked adult male* $\sim \mathcal{N}(\mu = 69, \sigma = 2.5)$
- $h =?$ *such that* $\mathrm{P}(X > h) = 0.04$
- $\mathrm{P}(X > h) = \mathrm{P}(X - \mu > h - 69) = \mathrm{P}(\frac{X-\mu}{\sigma} > \frac{h-69}{2.5}) = \mathrm{P}(Z > \frac{h-69}{2.5}) =$

0.04



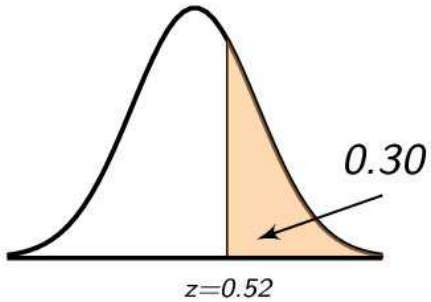
$$\frac{h-69}{2.5} = 1.75, h = 73.38 \text{ in}$$

Problem 4.6

The longevity of people living in a certain locality has a standard deviation of 14 years. What is the mean longevity if 30% of the people live longer than 75 years? Assume a normal distribution for life spans.

Solution

- $X =$ *the longevity of a randomly chosen person* $\sim \mathcal{N}(\mu = ?, \sigma = 14)$
- $\mu = ?$ *given that* $\mathrm{P}(X > 75) = 0.30$
- $\mathrm{P}(X > 75) = \mathrm{P}(Z > \frac{75-\mu}{14}) = 0.30$



$\frac{75-\mu}{14} = 0.52, \mu = 67.72$

Normal Approximation to Binomial Distribution

- $X = Binom(n, p)$
- $n =$ number of trials
- $p =$ probability of a success in one trial

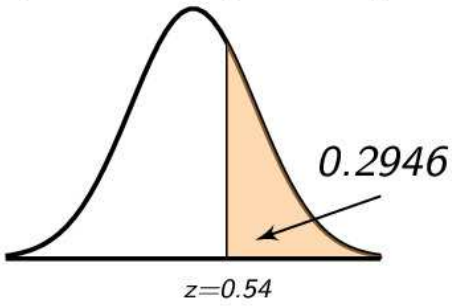
- $X = \mathcal{N}(\mu, \sigma)$
- $\mu = np$
- $\sigma = \sqrt{np(1-p)}$
- Approximation valid when $n > 40$, $np > 5$, and $n(1-p) > 5$

Problem 4.7

The unemployment rate in a certain city is 8.5%. A random sample of 100 people from the labor force is drawn. Find the approximate probability that the sample contains at least ten unemployed people.

Solution

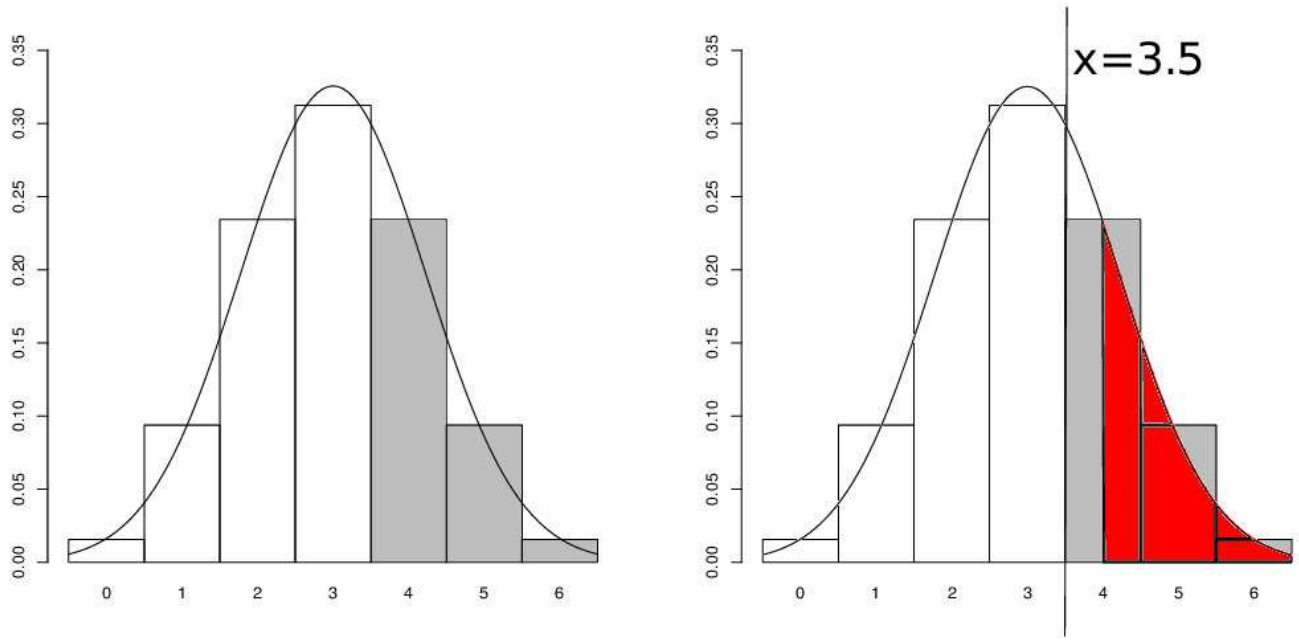
- $X =$ *the number of unemployed people* $\sim Bi(n = 100, p = 0.085)$
- $X \sim \mathcal{N}(\mu = np = 8.5, \sigma = \sqrt{np(1-p)} = 2.79)$
- $\mathrm{P}(X \geq 10) = \mathrm{P}(Z > \frac{10-8.5}{2.79}) = \mathrm{P}(Z > 0.54)$



$$\mathrm{P}(Z > 0.54) = 0.2946$$

Continuity Correction

The concept of continuity correction applies when a discrete distribution is approximated by a continuous distribution.



- Pr($X \geq 4$) by binomial distribution is gray
- Pr($X \geq 4$) by normal distribution is red
- Need to step 0.5 units left

Continuity Correction Rule

$$X = \text{discrete variable}$$

$$Y = \text{continuous variable}$$

Write integer inequality in two forms

$$P(X \leq k) = P(X < k + 1)$$

and take the average of the two borders, i.e.

$$P(X \leq k) = P(X < k + 1) \simeq P(Y < k + \frac{1}{2})$$

Expected Value and Variance

Random number = values with assigned probabilities

x	x_1	x_2	$\dots$	$\dots$	x_n
p	p_1	p_2	$\dots$	$\dots$	p_n

- Expected value $\mathrm{E}(X) = \sum_{i=1}^{n} x_i p_i$ = not a random number
- Expected value is a measure of center
- $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$
- Variance is a measure of spread
- Standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$

Good Properties of Expected Value and Variance

- $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
- $\mathrm{E}(cX) = c\,\mathrm{E}(X)$
- $\mathrm{E}(c) = c$
- If X and Y are independent then $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$

- $\mathrm{Var}(X) = \mathrm{E}(X^2) - E^2(X)$
- $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$
- If X and Y are independent then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ and $\sigma(X + Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}$
- Note that standard deviation follows Pythagorean rule $c = \sqrt{a^2 + b^2}$

Problem 4.8

The Attila Barbell Company makes bars for weight lifting. The weights of the bars are independent and are normally distributed with a mean of 720 ounces (45 pounds) and a standard deviation of 4 ounces. The bars are shipped 10 in a box to the retailers. The weights of the empty boxes are normally distributed with a mean of 320 ounces and a standard deviation of 8 ounces. The weights of the boxes filled with 10 bars are expected to be normally distributed with a mean of 7,520 ounces. What is the standard deviation?

Solution

- $X_i =$ *the weight of the i-th bar* $\sim \mathcal{N}(\mu = 720, \sigma = 4), i = 1 \dots 10$
- $Y =$ *the weight of the box* $\sim \mathcal{N}(\mu = 320, \sigma = 8)$
- $W =$ *the weight of the package* $= X_1 + \cdots + X_{10} + Y$
- $\mathrm{E}(W) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_{10}) + \mathrm{E}(Y) = 10 * 720 + 320 = 7,520$
- $\mathrm{Var}(W) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{10}) + Var(Y) = 10 * 4^2 + 8^2 = 224$
- $\sigma(W) = \sqrt{224} \simeq 15\ oz$
- *Note that* $\mathrm{Var}(10X + Y) = 10^2 * 4^2 + 8^2 = 1664 >> 224$

Sampling distribution

- Sample $X_1, X_2, \ldots, X_n$
- X_i are random numbers from certain population

Population = heights of adult males

Assume that X_i

- are from the same distribution
- are independent

X_1	X_2	X_3
176	181	190
181	190	176
190	176	181
164	176	188
190	190	190
...	...	...
...	...	...
...	...	...

Sample Mean

Sample mean is defined as $\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$

Assume that

- X_i are from the same distribution
  - $\mathrm{E}(X_i) = \mu$
  - $\mathrm{Var}(X_i) = \sigma^2$
- X_i and X_j are independent for $i \neq j$

$$\mathrm{E}(\overline{X}) = E\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) = \frac{1}{n}\mathrm{E}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{n\mu}{n} = \mu$$

The expected value of sample mean is the population mean

Law of Large Numbers

$$\overline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$$

$$\text{Var}(\overline{X}) = \text{Var}\left(\frac{X_1+X_2+\ldots+X_n}{n}\right) = \frac{1}{n^2}\text{Var}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{n\sigma^2}{n^2} = \sigma^2/n$$

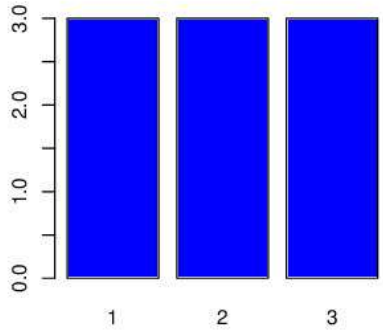
$$\text{Var}(\overline{X}) = \frac{\text{Var}(X)}{n} \text{ and } \sigma(\overline{X}) = \frac{\sigma(X)}{\sqrt{n}}$$

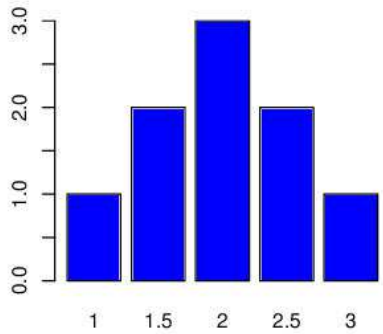
The larger the sample size, the less spread is the distribution of $\overline{X}$.

A Specific Example

X_1	X_2	$\overline{X}$
1	1	1
1	2	1.5
1	3	2
2	1	1.5
2	2	2
2	3	2.5
3	1	2
3	2	2.5
3	3	3

Population $= \{1, 2, 3\}$, sample size $n = 2$

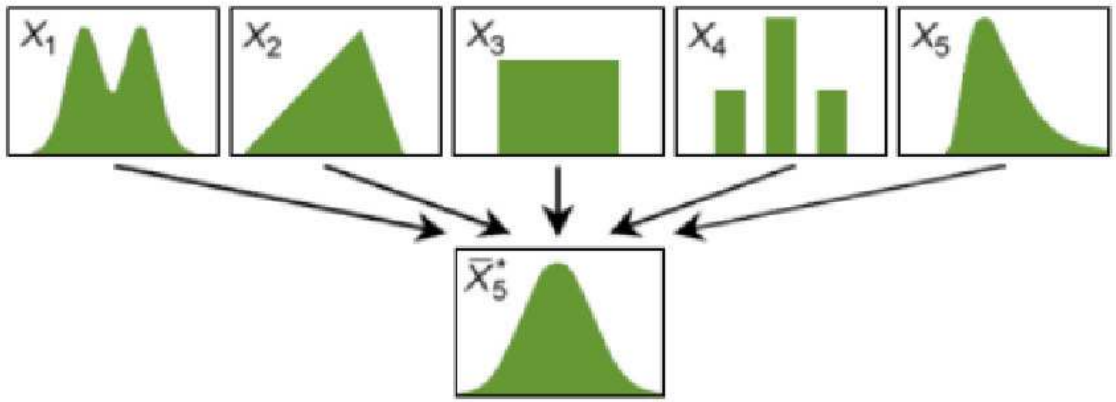




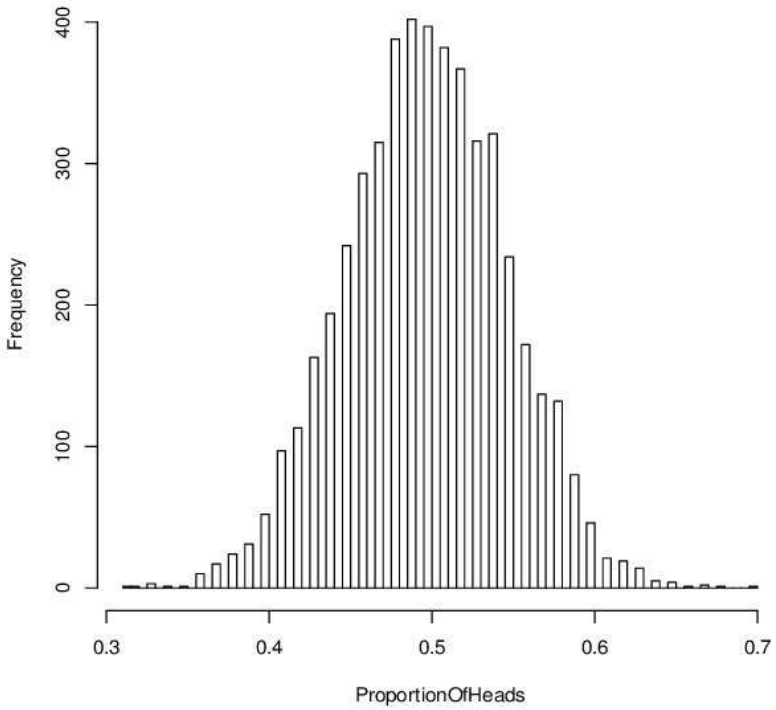
Central Limit Theorem

Theorem 1

The sum of a sufficiently large number of identically distributed independent random variables is approximately normally distributed regardless of the population distribution.



Normal Approximation to Binomial Distribution



X = number of successes in n trials

$$X = X_1 + X_2 + \ldots + X_n$$

$$X_i = \begin{cases} 0, & \text{if no success} \\ 1, & \text{if success} \end{cases}$$

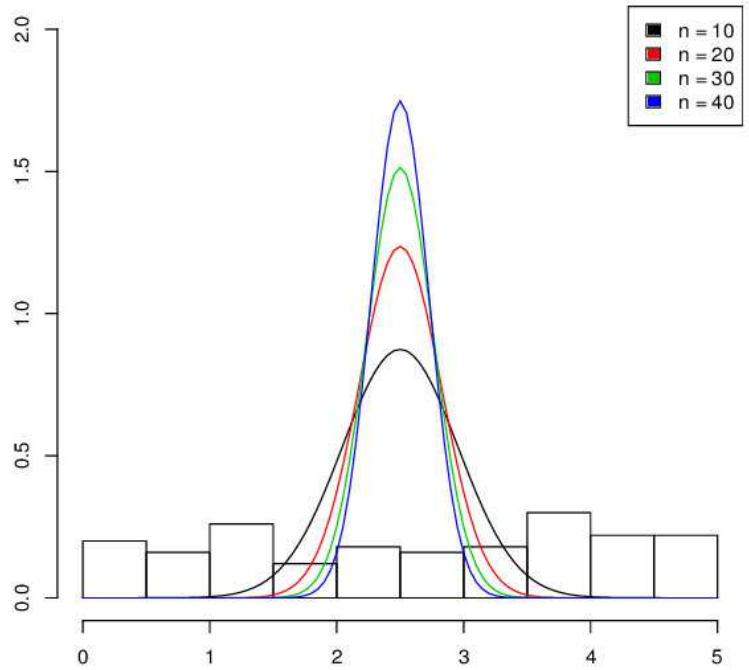
$$\mathrm{E}(X_i) = p$$

$$\mathrm{Var}(X_i) = p - p^2 = p(1-p)$$

$$\mathrm{E}(X) = np$$

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

Sampling distribution



$\overline{X}$ is approximately normal when $n > 40$

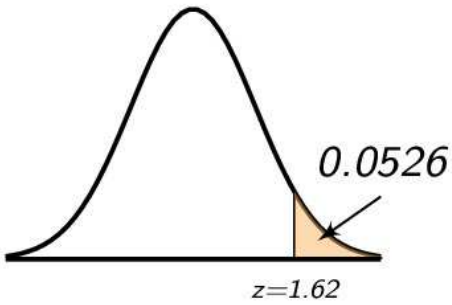
$\overline{X}$ is approximately normal *regardless* of the distribution of X

Problem 5.1

The average outstanding bill for delinquent customer accounts for a national department store chain is \$187.50 with a standard deviation of \$54.50. In a simple random sample of 50 delinquent accounts, what is the probability that the mean outstanding bill is over \$200?

Solution

- $X =$ *random outstanding bill for a delinquent customer*
- $\mathrm{E}(X) = \mu_X = 187.50, \quad \sigma_X = 54.50$
- $\overline{X} \sim \mathcal{N}(\mu = 187.50, \sigma = \frac{54.50}{\sqrt{50}})$
- $\mathrm{P}(\overline{X} > 200) = \mathrm{P}(Z > \frac{200-187.50}{54.50/\sqrt{50}}) = \mathrm{P}(Z > 1.62)$



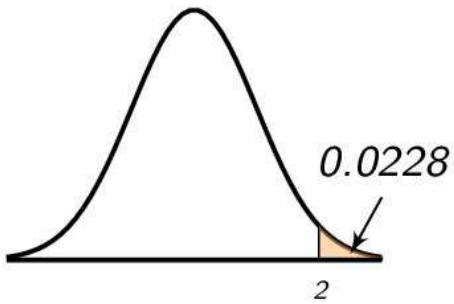
$$\mathrm{P}(Z > 1.62) = 0.0526$$

Problem 5.2

A summer resort rents rowboats to customers but does not allow more than four people to a boat. Each boat is designed to hold no more than 800 pounds. Suppose the distribution of adult males who rent boats, including their clothes and gear, is normal with a mean of 190 pounds and standard deviation of 10 pounds. If the weights of individual passengers are independent, what is the probability that a group of four adult male passengers will exceed the acceptable weight limit of 800 pounds?

Solution

- $X = \textit{weight of a passenger} \sim \mathcal{N}(\mu = 190, \sigma = 10)$
- $\mathrm{P}(X_1 + X_2 + X_3 + X_4 > 800) = \mathrm{P}(\overline{X} > 200) = ?$
- $\mathrm{P}(\overline{X} > 200) = \mathrm{P}(Z > \frac{200-190}{10/\sqrt{4}}) = \mathrm{P}(Z > 2) = 0.0228$



Population normality

- If n is large then $\overline{X}$ is approximately normal regardless of the population distribution
- A non-trivial linear combination of independent normal distributions is normal
- If n is small AND population is normal then $\overline{X}$ is also normal
- For large n, population normality is **not** required
- For small n, population normality **is** required

Summary

- Probability is a non-negative measure normalized to 1
- Independent events are such that $\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$
- Discrete distributions take only isolated values, while continuous distributions also take all intermediate values
- Continuity correction is needed when approximating a discrete distribution by a continuous distribution
- Expected value is a linear operation, i.e. $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$
- Standard deviation is not, namely if X and Y are independent then $\sigma(X + Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}$ (Pythagorean theorem)
- Sampling distribution if the distribution of sample means; its mean is the same as in the population, but σ is $\sqrt{n}$ times smaller (The Law of Large Numbers)
- Sample mean is approximately normally distributed when n is large