Week 12: Central limit theorem (cont), descriptive statistics

Armenak Petrosyan

Last time

Random samples

- For i.i.d. random variables, we say $X_1, \ldots, X_n$ form a **random sample of size** n from the common distribution.
- We use the values $X_1(s), \ldots, X_n(s)$ to model random samples taken during a single run of the experiment.

Example

The laboratory assistant catches (randomly samples) n insects of the same type during the experiment and measures their wing lengths. The corresponding lengths will be $X_1(s), \ldots, X_n(s)$.
If another assistant at a different laboratory does the same experiment, his measurement may be $X_1(s'), \ldots, X_n(s')$ for potentially different from s value of s'.

Example

We randomly pick n number of students from Georgia Tech and measure their heights.

- $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2.
- μ is called **population mean** and σ^2 is called **population variance**.

$$\boxed{\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}} \quad \textbf{(sample mean)}.$$

- $E[\bar{X}_n] = \mu$.
- $\mathrm{Var}(\bar{X}_n) = \frac{1}{n}\sigma^2$.

Theorem (Strong law of large numbers)

Let $X_1, \dots, X_n$ be a random sample (i.i.d. random variables) with mean μ then

$$P(\lim_{n \to \infty} \bar{X}_n = \mu) = 1.$$

Example

If we throw a die and record the values, the sample mean will converge to 3.5 with probability 1.

$$\boxed{\bar{Z}_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}$$

We showed that

$$M_{Z_n}(t) \to e^{\frac{t^2}{2}}, \quad n \to \infty.$$

In conclusion:

The mgf of $\bar{Z}_n$ converges to the mgf of $N(0,1)$.

Tuesday class

Convergence in distribution

Definition

Let W_n, W be any random variables. We say that W_n **converges in distribution** to W if

$$\lim_{n\to\infty} F_{W_n}(w) = F_W(w), \text{ for every } w \in \mathbb{R}.$$

- W_n converges to W in distribution is equivalent to

$$\lim_{n\to\infty} P(a < W_n \le b) = P(a < W \le b), \text{ for every } a, b \in \mathbb{R}.$$

- Convergence of mgf-s implies convergence in distribution.

Fact

Let $\delta > 0$. If

$$\lim_{n\to\infty} M_{W_n}(t) = M_W(t) \quad \text{for every } |t| < \delta$$

then $W_n \to W$ in distribution.

Theorem (Central limit theorem)

Assume $X_1, \ldots, X_n$ are i.i.d. for which μ and σ^2 exists, and let $\bar{Z}_n$ be the Z-score of $\bar{X}_n$. Then $\bar{Z}_n$ converges to $N(0,1)$ in distribution: for every $a, b \in \mathbb{R}$,

$$\lim_{n\to\infty} P(a < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t$$

or, equivalently,

$$\lim_{n\to\infty} P(\mu + \frac{a}{\sqrt{n}}\sigma < \frac{1}{n}\sum_{i=1}^{n} X_n \leq \mu + \frac{b}{\sqrt{n}}\sigma) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t.$$

- The CLT claims that the cdf of Z_n is close to the cdf of $N(0,1)$ when n is large.
- After the linear transform, it says that the sample mean $\bar{X} = \frac{X_1 + \cdots X_n}{n}$ is approximately $N(\mu, \frac{\sigma^2}{n})$.
- When X_i-s are normal, $\bar{X}$ is exactly $N(\mu, \frac{\sigma^2}{n})$.

Experiment: data collection

Data collection:

- We take i.i.d. random variables $X_1, \ldots, X_n$ from
  1. Uniform distribution on $[2, 4]$.
  2. Exponential distribution with $\theta = 2$.
- Then
  1. $\mu = \frac{4+2}{2} = 3, \quad \sigma^2 = \frac{(4-2)^2}{12} = \frac{1}{3}$
  2. $\mu = \theta, \quad \sigma^2 = \theta^2.$
- We sample values $X_1, \ldots, X_n$ and record the sample mean.
- We compute 1000 sample means this way.
- We compute the Z-score of the means.
- Our data consists of 1000 Z-score values.

Experiment: density histogram

Since this is a continuous-type data (collected from a continuous distribution), we will represent the distribution of data by grouping them into classes and computing the relative frequency histogram on these classes.

- Determine the largest and smallest value of the data - $z_{\min}, z_{\max}$.
- Divide the interval into equally sized intervals called **class intervals**

$$(c_0, c_1], (c_1, c_2], \ldots, (c_{14}, c_{15}].$$

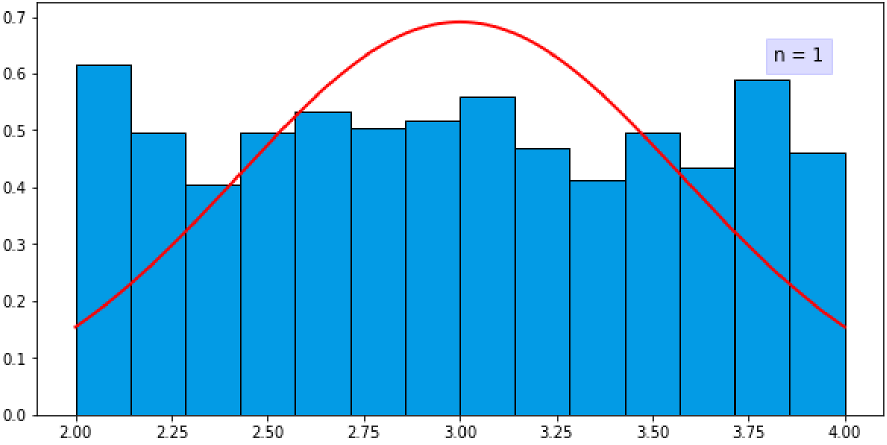
- The density on each interval is computed by

$$\frac{\text{number of data in the interval}}{n \times \text{interval length}}.$$

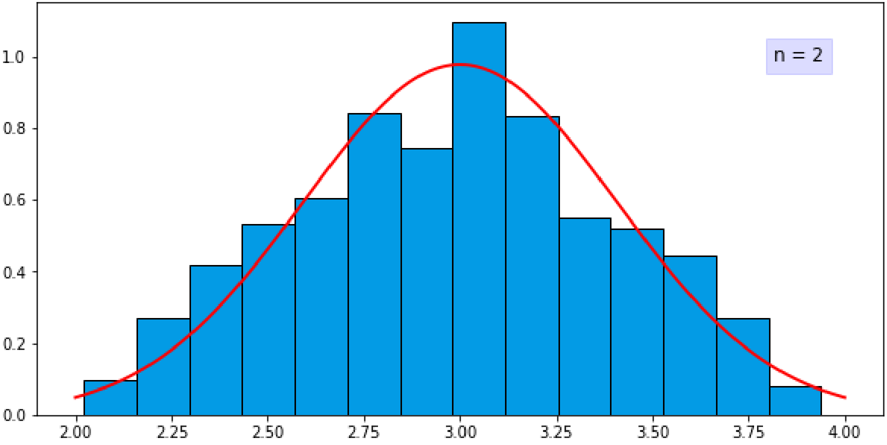
(we divide by the interval length because we are looking at the pdf-s).

- We plot the density histogram for different values of n.
- We also plot the standard normal distribution.

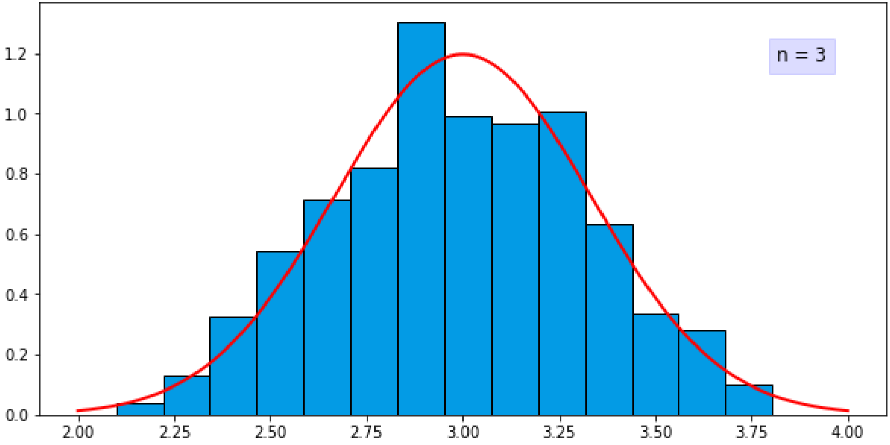
Uniform



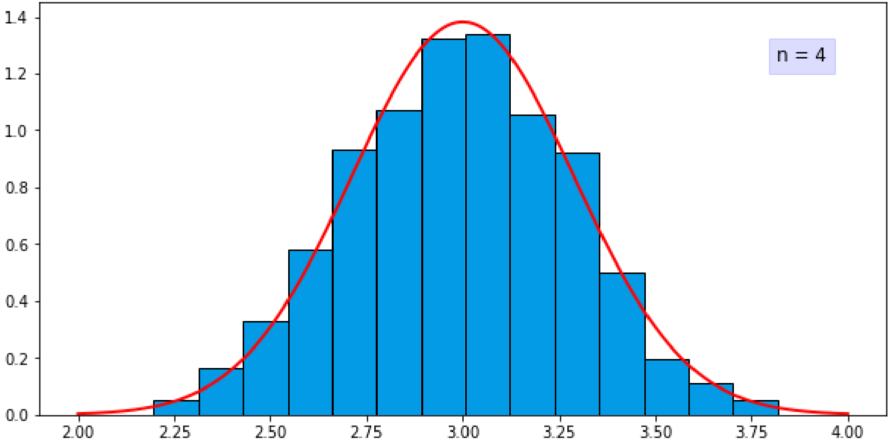
Uniform



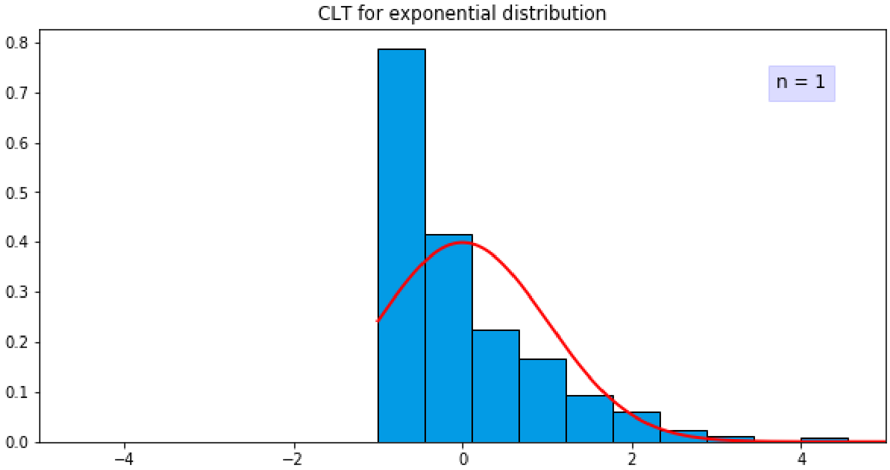
Uniform



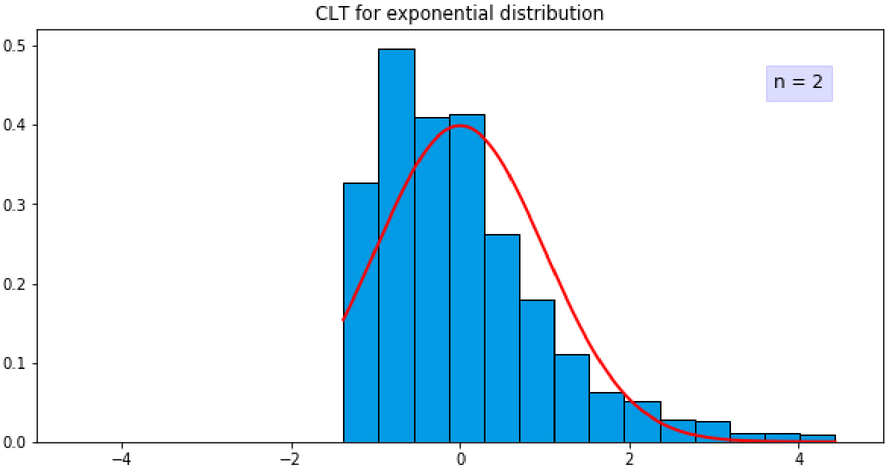
Uniform



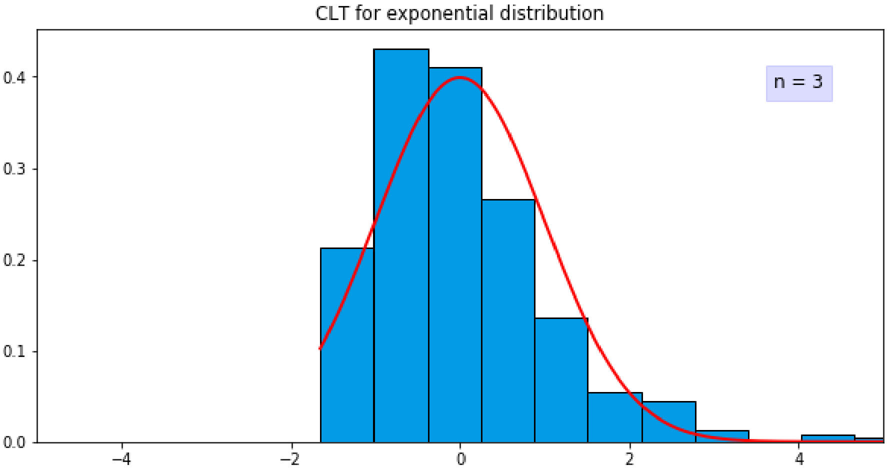
Exponential

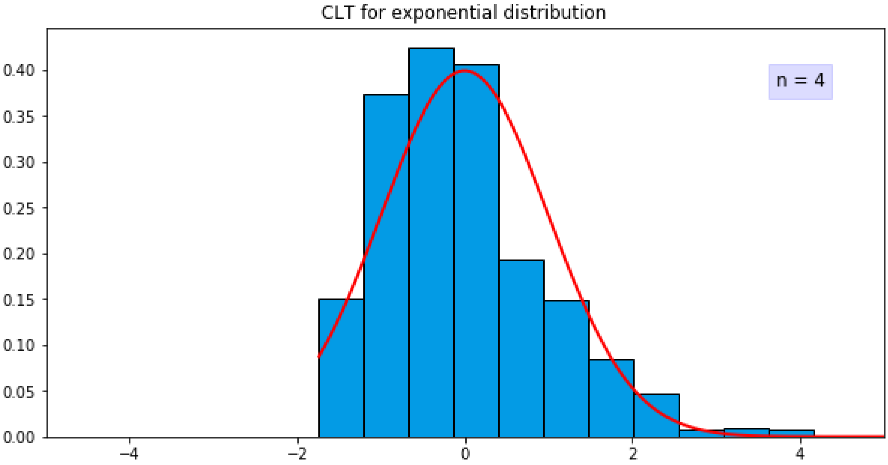


Exponential

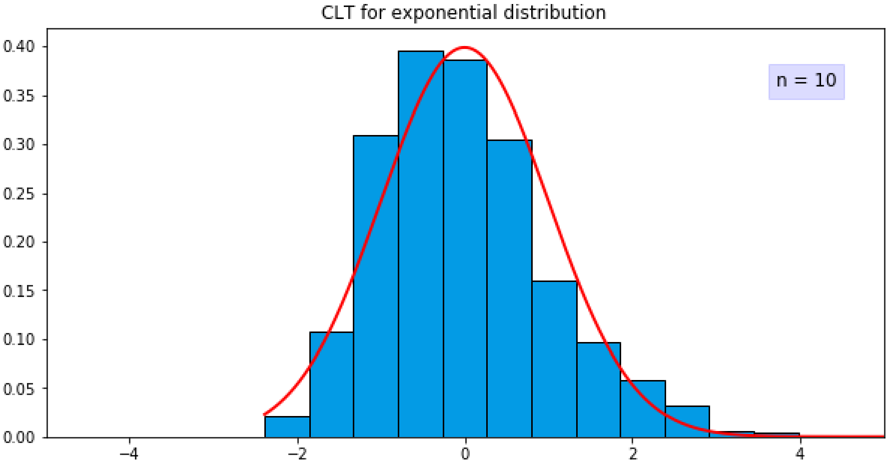


Exponential

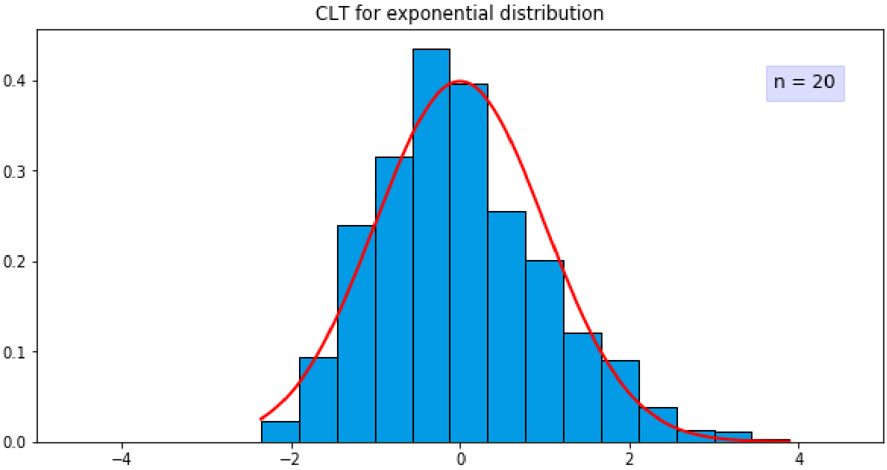



CLT for exponential distribution
n = 4
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4

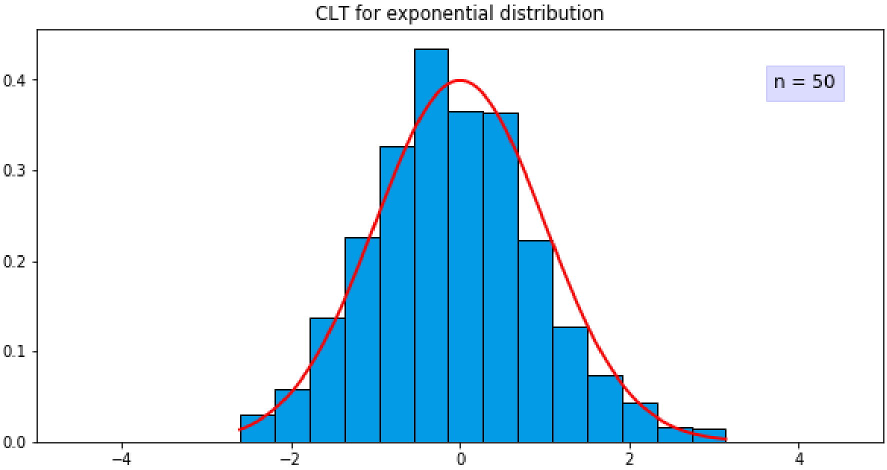
Exponential



Exponential



Exponential



Exercise 1

Problem (5.6-13 in the textbook)

The tensile strength X of paper, in pounds per square inch, has $\mu = 30$ and $\sigma = 3$. A random sample of size $n = 100$ is taken from the distribution of tensile strengths. Compute the probability that the sample mean $\bar{X}$ is greater than 29.5 pounds per square inch.

Solution

- *Let* $\bar{Z} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{10(\bar{X}-30)}{3}$.
- $$\begin{aligned} P(\bar{X} \geq 29.5) &= P(\bar{Z} \geq \frac{10(\bar{2}9.5 - 30)}{3}) \\ &\approx P(\bar{Z} \geq -1.67) \\ &\approx P(\bar{Z} \leq 1.67) \\ &\boxed{\approx 0.9525}. \end{aligned}$$

Example: basic Brownian motion

- Particle starts at point 0 on the line.
- Particle moves randomly left or right by the increment of X_i at time i.
- We assume X_i are i.i.d. with $\mu = 0$, $\sigma = 1$.
- Y_n is position of the particle at time n:

$$Y_n = \sum_{i=1}^{n} X_i.$$

- When n is sufficiently large

$$Y_n = n\bar{X}_n = n(\frac{\sigma}{\sqrt{n}}\bar{Z}_n + \mu) = \sqrt{n}\bar{Z}_n$$

  and so it approximately has the distribution $N(0, n)$.
- If we simulate an experiment with large number of particles, they will create a cloud that has approximately normal distribution around 0 and radius (expansion speed) $\sqrt{n}$.
- We can do the same in 3D applying CLT to each coordinate (ignoring gravity) .
- If you drop ink into water, the spread of the ink over time looks like normal distribution.

Thursday class

Normal approximation to binomial distribution

- The central limit theorem holds true both for continuous and discrete random variables.
- Let X_i be i.i.d. with Bernoulli distribution and $P(X_i = 1) = p$.
- $\mu = p, \quad \sigma^2 = p(1-p)$.
- Notice that
$$Y = \sum_{i=1}^{n} X_i$$
has binomial distribution with parameters (n, p).
- From CLT,
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{Y - np}{\sqrt{np(1-p)}}$$
is close to $N(0, 1)$.
- More precisely,
$$\lim_{n \to \infty} P(a < \frac{Y - np}{\sqrt{np(1-p)}} \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, \mathrm{d}t.$$
- To compute $P(A < Y \le B)$, we do change of variables
$$P(A < Y \le B) = P(\frac{A - np}{\sqrt{np(1-p)}} < \frac{Y - np}{\sqrt{np(1-p)}} \le \frac{B - np}{\sqrt{np(1-p)}}).$$

Half correction

- Since Y is discrete and the normal distribution is continuous, to approximate the pmf $P(Y = k)$, we use instead
$$P(k - \frac{1}{2} < Y \leq k + \frac{1}{2}).$$
This is called **half correction**.
- Similarly, we do (by adding the approximate pmf-s)
$$P(k_1 \leq Y < k_2) \approx P(k_1 - \frac{1}{2} < Y \leq k_2 - \frac{1}{2}).$$

Exercise 2

Problem (5.7-15)

In the United States, the probability that a child dies in his or her first year of life is about $p = 0.01$. (It is actually slightly less than this.) Consider a group of 5000 such infants. What is the probability that between 45 and 53, inclusive, die in the first year of life?

Solution

- $n = 5000, \quad p = 0.01, \quad \sigma^2 = p(1-p)$
- *Let* $F(x)$ *be cdf of* $N(0,1)$. *We want to estimate*

$$\begin{aligned}
P(45 \le Y \le 53) &\approx P(44.5 < Y \le 53.5) \\
&= P\left(\frac{44.5 - np}{\sqrt{np(1-p)}} < \frac{Y - np}{\sqrt{np(1-p)}} \le \frac{53.5 - np}{\sqrt{np(1-p)}}\right) \\
&= P\left(\frac{44.5 - 50}{\sqrt{49.5}} < \frac{Y - np}{\sqrt{np(1-p)}} \le \frac{53.5 - 50}{\sqrt{49.5}}\right) \\
&= P\left(-0.8 < \frac{Y - np}{\sqrt{np(1-p)}} \le 0.5\right) \\
&\approx F(0.5) - F(-0.8) = F(0.5) - (1 - F(0.8)) \\
&= 0.6915 - (1 - 0.7881) \boxed{= 0.4796}.
\end{aligned}$$

Normal approximation to Poisson distribution

- Let $X_1, \dots, X_n$ be i.i.d. Poisson with rate λ_0.
- It can be checked, that
$$Y = \sum_{i=1}^{n} X_i$$
again has Poisson distribution but with rate $\lambda = n\lambda_0$.
- From CLT,
$$\bar{Z} = \frac{\bar{X} - \lambda_0}{\frac{\sqrt{\lambda_0}}{\sqrt{n}}} = \frac{\frac{Y}{n} - \lambda_0}{\frac{\sqrt{\lambda_0}}{\sqrt{n}}} = \frac{Y - \lambda}{\sqrt{\lambda}}$$
is approximately $N(0,1)$ when n is large, or equivalently, λ is large.

If Y has Poisson distribution with rate λ then
$$\bar{Z} = \frac{Y - \lambda}{\sqrt{\lambda}}$$
is approximately $N(0,1)$ when λ is large.

Normal vs Binomial vs Poisson

- If n is large while $\lambda = np$ is small , then (we have discussed this, see week 4 slides)

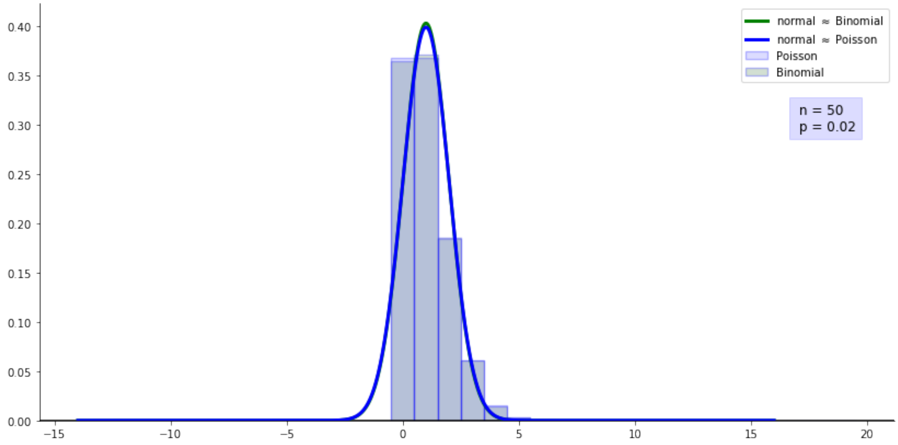
  Binomial(n,p) $\approx$ Poisson(np).

- When np and $n(1-p)$ are large,

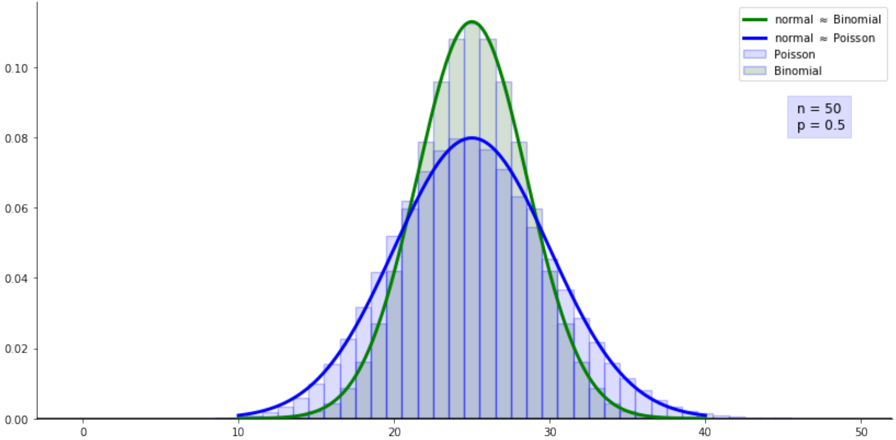
  Binomial(n,p) $\approx$ Normal (np, np(1-p)).

- When λ is large

  Poisson$(\lambda) \approx$ Normal(λ, λ).



n is large but $np = 1$ is not.



$np = 25$ is large so binomial and Poisson are different but close to corresponding normal approximations.

Descriptive statistics

Here, **data** is a sequence of any univariate numerical measurements $\{x_1, \ldots, x_n\}$.

- Can be any type of data; does not have to be sampled from a distribution.

Descriptive statistics are a collection of data summaries providing information about the data.

- **Mean** of data (measures central tendency):

$$\mu = \frac{x_1 + \cdots + x_n}{n}.$$

- **Variance** of data (measures dispersion):

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}.$$

- The **five-number summary**.
- **Interquartile range**.
- Etc.

The process of analyzing data using its descriptive statistics is called **exploratory data analysis**.

Order statistics

Definition

If the data is reordered in ascending order,

$$\{x_1, \ldots, x_n\} = \{y_1, \ldots, y_n\}$$

with $y_1 \leq y_2 \leq \cdots \leq y_n$ then y_i is called the **i-th order statistic** of the data.

- The first order statistic is the minimum and the n-th order statistic is the maximum.
- If x_i is the r-th order statistic then we say x_i has rank r.

Example

In $\{0.5, 0.3, 05\}$ the corresponding ranks are $2, 1, 3$.

Definition

For $0 < p < 1$, the $100p$-th percentile of the data is called the following number $\pi_p \in \mathbb{R}$. Write $(n+1)p = r + \alpha$ where r is an integer (it is the $\lfloor (n+1)p \rfloor$) and $0 \le \alpha < 1$. Take

$$\pi_p = y_r + \alpha(y_{r+1} - y_r)$$

where y_r and y_{r+1} are the r and $r+1$ order statistics.

- Approximately $100p\%$ of data values are smaller than $\tilde{\pi}_i$ (np data values) and the rest (approximately $n(1-p)$ values) are greater.

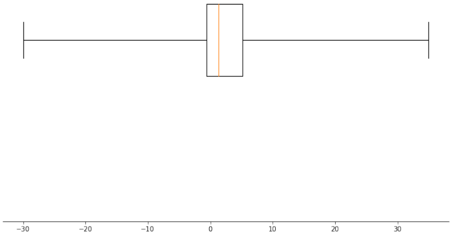
Definition

$\pi_{0.25}, \pi_{0.5}, \pi_{0.75}$ are called **first, second and third quartiles** and denoted by q_1, q_2, q_3 correspondingly.

- The second quartile is also called **median**.
- The minimum, maximum and the three quartiles are called the **five-number summary** of the data.
- $q_3 - q_1$ is called **interquartile range** or **IQR**.

Box plot

- Box plot is a box with two whiskers.
- The left and right corners are at q_1 and q_3.
- There is a vertical segment drawn at the median. Position of this line shows the **skewness** of the data.
- The left whisker extends from the minimum to q_1 and the right whisker extends from q_3 to the maximum.



Outliers

Outliers are the atypical elements of the data set.

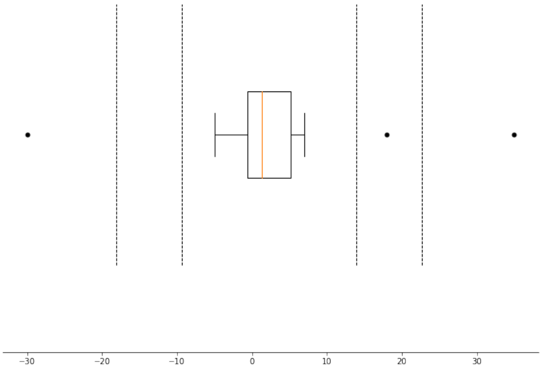
Example

The majority of university professors in the department have 50,000$-150,000$ annual income. Professor M is an atypical professor; he has 500,000$ annual income.

- The data values 1.5 IQR distance away from the left and right sides (i.e. outside the interval $[q_1 - 1.5 \cdot \text{IQR}, q_1 + 1.5 \cdot \text{IQR}]$) are called **suspected outliers**.
- The data values 3 or more IQR distance away from the left and right sides are called **outliers**.
- We draw vertical lines at points

$$q_1 - 1.5 \cdot \text{IQR}, \quad q_3 + 1.5 \cdot \text{IQR}, \quad q_1 - 3 \cdot \text{IQR}, \quad q_3 + 3 \cdot \text{IQR}.$$

- The inner pair and the outer pair are called correspondingly **Tukey's inner and outer fences**.
- We extend the whiskers up to the left-most and right-most data points inside the inner fence.
- Outliers are marker with black circles.


−30
−20
−10
0
10
20
30

- Both mean and median measure central tendency.
- Both variance and interquartile range measure dispersion.
- Median and interquartile range are more robust to outliers: an outlier does not affect them too dramatically.

Exercise 3

Problem

Let the data be given as

$$\{40, 20, -5, 10, -30, 13, 500, 9\}.$$

(a) *Find the mean.*

(b) *Compute the 5-number summary of the data.*

(c) *Draw the box plot with Tukey's fences and outliers.*

Solution

(a)

$$\mu = \frac{40 + 20 + (-5) + 10 + (-30) + 13 + 500 + 9}{8} \approx 69.625.$$

(b) *Rearrange in ascending order*

$$\{-30, -5, 9, 10, 13, 20, 40, 500\}.$$

Solution (cont.)

- $\min = -30, \quad \max = 500.$
- $(n+1)0.25 = 9 \cdot 0.25 = 2.25 = 2 + 0.25.$ *And so*

$$q_1 = \pi_{0.25} = y_2 + 0.25 \cdot (y_3 - y_2) = -5 + 0.25 \cdot (9 - (-5)) = -1.5.$$

- $(n+1)0.5 = 9 \cdot 0.5 = 4.5 = 4 + 0.5.$ *And so*

$$q_2 = \pi_{0.5} = y_4 + 0.5 \cdot (y_5 - y_4) = 10 + 0.5 \cdot (13 - 10) = 11.5.$$

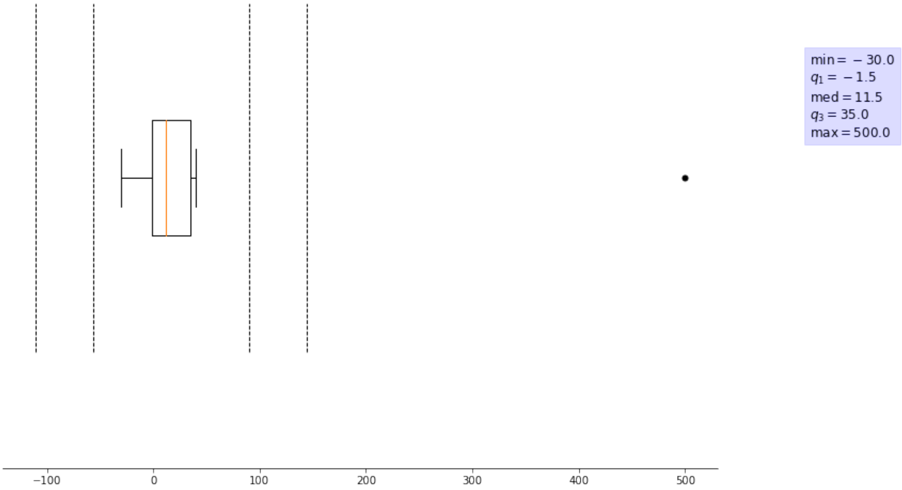
- $(n+1)0.75 = 9 \cdot 0.75 = 6.75 = 6 + 0.75.$ *And so*

$$q_3 = \pi_{0.75} = y_6 + 0.75 \cdot (y_7 - y_6) = 20 + 0.75 \cdot (40 - 20) = 35.$$

- $\mathit{IQR} = q_3 - q_1 = 36.5.$
- *Inner fences*

$$\begin{aligned} q_1 - 1.5 \cdot \mathit{IQR} &= -1.5 - 1.5 \cdot 36.5 = -56.25 \\ q_3 + 1.5 \cdot \mathit{IQR} &= 35 + 1.5 \cdot 36.5 = 89.75 \\ q_1 - 3 \cdot \mathit{IQR} &= -1.5 - 3 \cdot 36.5 = -111 \\ q_3 + 3 \cdot \mathit{IQR} &= 35 + 3 \cdot 36.5 = 144.5 \end{aligned}$$

- *500 is an outiler.*


$\min = -30.0$
$q_1 = -1.5$
$\text{med} = 11.5$
$q_3 = 35.0$
$\max = 500.0$
−100
0
100
200
300
400
500

Warning

In different software, the percentile is computed differently:

$$\pi_p = y_r + \alpha(y_{r+1} - y_r)$$

where

- in Excel: $r = \lfloor (n+1)p \rfloor$ and $\alpha = (n+1)p - r$ **(this is how we defined)**
- in Numpy: $r = \lfloor (n-1)p + 1 \rfloor$ and $\alpha = (n-1)p + 1 - r$

You may get different results based on what you use. Make sure it matches our definition.