Statistics for Computing MA4413

Lecture 12

The Central Limit Theorem

Kevin Burke

kevin.burke@ul.ie

Distributions Studied

We have studied the most commonly used distributions - many other possible distributions exist.

Distribution	Variable Type	$E(X)$	$Var(X)$	$Sd(X)$
Bernoulli	Categorical: $x \in \{0, 1\}$	p	$p(1-p)$	$\sqrt{p(1-p)}$
Binomial	Discrete: $x \in \{0, 1, \ldots, n\}$	np	$np(1-p)$	$\sqrt{np(1-p)}$
Poisson	Discrete: $x \in \{0, 1, \ldots, \infty\}$	λ	λ	$\sqrt{\lambda}$
Exponential	Continuous: $t \in [0, \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{1}{\lambda}$
Normal	Continuous: $x \in (-\infty, \infty)$	μ	σ^2	σ

Mean of Random Variables

We are now interested in the **mean** of a *sample of independent random variables*.

Let $X_1, X_2, \ldots, X_n$ be a sample of *numeric* random variables which come from the *same* probability distribution. The mean of this sample:

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

is *also a random variable* - it varies from sample to sample.
(capital $\bar{X}$ refers to the random variable and small $\bar{x}$ refers to a specific value)

In particular we are interested in the *distribution* of $\bar{X}$ so that we can determine how much it varies from sample to sample.

Note on Bernoulli Variables

For *numeric* variables (e.g., binomial, Poisson, exponential, normal), we are interested in the mean $\overline{X}$.

For *categorical* Bernoulli variables, where $x \in \{0, 1\}$, we have

$$\frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\text{the number of 1s}}{n} = \widehat{P}.$$

This **proportion**, $\widehat{P}$, is a random variable - it varies from sample to sample. (random variable $= \widehat{P}$, specific value $= \hat{p}$)

We can see that $\widehat{P}$ is “the mean for categorical variables” but, to avoid confusion, we will *never* use $\overline{X}$ here.

Notation

We began with statistics, using the symbols μ and σ to denote the *true* (unknown) mean and standard deviation (for a *numeric* data).

In the *theoretical realm* of probability we have *calculated* these true values. Notation: $E(X)$ and $Sd(X)$.

We now return to using the symbols μ and σ as we will soon be returning to statistical matters. (next lecture)

Note: for *categorical* data we have the true *proportion*, p.

The Central Limit Theorem

The **central limit theorem**, or CLT, is a fundamental result in statistical theory. It states the following:

Regardless of the distribution of $X_1, \ldots, X_n$, the sample mean $\overline{X}$ has a *normal distribution* when the sample size, n, is large (also holds for $\widehat{P}$).

This fact allows us to test statistical hypotheses about the true mean based on a sample of data (more on this next lecture).

The theorem is “central” as is the core of most statistical theory and also it refers to the mean - a measure of centrality.

Ubiquity of the Normal Distribution

The central limit theorem also provides one explanation for the ubiquity of the normal distribution in practice.

Various quantities can be viewed as the average result of many other variables which leads to a normality, e.g., the weight or height of an animal is the net effect of biological and environmental variables.

The Central Limit Theorem: Result

For a sample of **independent variables**, $X_1, X_2, \ldots, X_n$, which come from a distribution with mean, $\mu = E(X)$, and standard deviation, $\sigma = Sd(X)$, the result of the central limit theorem is that

$$\boxed{\bar{X} \sim \text{Normal}\left(\mu,\ \frac{\sigma}{\sqrt{n}}\right)}$$

when n is large.

In practice, this typically works well as along as $\boxed{n > 30}$.

Standard Error

The standard deviation of $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$. However, to avoid confusion, we won't use the phrase "standard deviation" in this setting.

In the interest of clarity:

- Call $\frac{\sigma}{\sqrt{n}}$ the **standard error** of $\overline{X}$.
- Let $\boldsymbol{\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}}$ denote this quantity.

Do not mix up:

- The standard deviation: σ describes how *individual values*, $X_1, \ldots, X_n$, vary around μ.
- The standard error: $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ describes how *sample means* (from different samples) vary around μ.

Reducing Standard Error

It is clear from the formula for standard error:

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

that **increasing the sample size** reduces the standard error.

This reflects the fact that, in larger samples, $\overline{X}$, will be closer to the true mean, μ.

Examples

$$X_1, \ldots, X_n \sim \text{Bernoulli}(p) \qquad \Rightarrow \widehat{P} \sim \text{Normal}\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

$$X_1, \ldots, X_n \sim \text{Binomial}(n, p) \qquad \Rightarrow \bar{X} \sim \text{Normal}\left(n\,p, \frac{\sqrt{np(1-p)}}{\sqrt{n}}\right)$$

$$X_1, \ldots, X_n \sim \text{Poisson}(\lambda) \qquad \Rightarrow \bar{X} \sim \text{Normal}\left(\lambda, \frac{\sqrt{\lambda}}{\sqrt{n}}\right)$$

Examples

$$T_1, \ldots, T_n \sim \text{Exponential}(\lambda) \qquad \Rightarrow \bar{T} \sim \text{Normal}\left(\frac{1}{\lambda}, \frac{\frac{1}{\lambda}}{\sqrt{n}}\right)$$

$$X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma) \qquad \Rightarrow \bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Examples: Light Bulbs

Let's assume that the lifetime of a light bulb is $T \sim$ Exponential with $E(T) = 1000$ hours $\Rightarrow \lambda = \frac{1}{E(T)} = \frac{1}{1000} = 0.001$.

Of course, we can answer probability questions pertaining to the life time of *one* bulb using the probability function

$$\Pr(T > t) = e^{-\lambda t} = e^{-0.001t}.$$

However, for the *mean life* in a sample of n bulbs, we need to use the central limit theorem:

$$\bar{T} \sim \text{Normal}\left(\frac{1}{\lambda}, \frac{\frac{1}{\lambda}}{\sqrt{n}}\right) = \text{Normal}\left(1000, \frac{1000}{\sqrt{n}}\right).$$

Examples: Light Bulbs

Let’s say we look at a sample of 25 bulbs. The mean life for this sample is

$$\bar{T} \sim \text{Normal}\left(1000, \frac{1000}{\sqrt{25}}\right) = \text{Normal}(1000, 200).$$

What is the probability that the sample mean is greater than 1200 hours?

$$\begin{aligned}\Pr(\bar{T} > 1200) &= \Pr(Z > \tfrac{1200-1000}{200}) \\ &= \Pr(Z > 1) \\ &= 0.1587.\end{aligned}$$

Examples: Light Bulbs

We may also wish to calculate the 95% limits:

$$\mu \pm z_{0.025}\,\sigma(\overline{T})$$

$$1000 \pm 1.96\,(200)$$

$$1000 \pm 392$$

$$\Rightarrow [608, 1392].$$

Thus, 95% of the time the mean (for a sample of 25 bulbs) will be contained in the interval [608, 1392].

Question 1

Let's assume that individuals' test scores are
$X \sim \text{Normal}(\mu = 60, \sigma = 10)$.

a) For a sample of size n, what is the distribution of $\bar{X}$?

b) In a sample of 30 individuals, what is $\Pr(\bar{X} > 66)$?

c) Calculate 99% limits for a group of 30 individuals.

d) Calculate 99% limits for a group of 50 individuals.

Difference Between Two Means

Often we are interested in the **difference in the means** for two groups.

We know, by the CLT, that

$$\overline{X}_1 \sim \text{Normal}\left(\mu_1,\ \frac{\sigma_1}{\sqrt{n_1}}\right) \qquad \text{and} \qquad \overline{X}_2 \sim \text{Normal}\left(\mu_2,\ \frac{\sigma_2}{\sqrt{n_2}}\right).$$

We also know, from Lecture11, how to deal with the difference between two normal variables. In particular the standard error is

$$\sigma(\overline{X}_1 - \overline{X}_2) = \sqrt{[\sigma(\overline{X}_1)]^2 + [\sigma(\overline{X}_2)]^2}.$$

$$\Rightarrow \boxed{\overline{X}_1 - \overline{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_2,\ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_1}}\right)}$$

Example: Errors

Let’s assume that the number of errors made by two machines in a given day is as follows:

$$X_1 \sim \text{Poisson}(\lambda_1 = 10) \qquad X_2 \sim \text{Poisson}(\lambda_1 = 8)$$

$$\Rightarrow \mu_1 = \lambda_1 = 10 \qquad \Rightarrow \mu_2 = \lambda_2 = 8$$

$$\Rightarrow \sigma_1^2 = \lambda_1 = 10 \qquad \Rightarrow \sigma_2^2 = \lambda_2 = 8$$

We record the number of errors produced by each machine on 40 separate days and compute the sample means. The difference is:

$$\bar{X}_1 - \bar{X}_2 \sim \text{Normal}\left(10 - 8 = 2,\ \sqrt{\frac{10}{40} + \frac{8}{40}} \approx 0.671\right)$$

Example: Errors

What is the probability that the difference in means is more than 3?

$$\begin{aligned}\Pr(\bar{X}_1 - \bar{X}_2 > 3) &= \Pr(Z > \tfrac{3-2}{0.671}) \\ &= \Pr(Z > 1.49) \\ &= 0.0681.\end{aligned}$$

Compute the interval within which the difference lies 95% of the time.

$$(\mu_1 - \mu_2) \pm z_{0.025}\, \sigma(\bar{X}_1 - \bar{X}_2)$$

$$2 \pm 1.96\,(0.671)$$

$$2 \pm 1.315$$

$$\Rightarrow [0.685, 3.315].$$

Difference Between Two Proportions

For categorical data we calculate the **difference in proportions**.

By the CLT we have that

$$\widehat{P}_1 \sim \text{Normal}\left(p_1,\ \sqrt{\frac{p_1(1-p_1)}{n_1}}\right),$$

$$\widehat{P}_2 \sim \text{Normal}\left(p_2,\ \sqrt{\frac{p_2(1-p_2)}{n_2}}\right).$$

$$\Rightarrow \boxed{\widehat{P}_1 - \widehat{P}_2 \sim \text{Normal}\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)}$$

Summary

		Standard Error
One Mean	$\bar{X}$	$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
One Proportion	$\widehat{P}$	$\sigma(\widehat{P}) = \sqrt{\frac{p\,(1-p)}{n}}$
Two Means	$\bar{X}_1 - \bar{X}_2$	$\sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Two Proportions	$\widehat{P}_1 - \widehat{P}_2$	$\sigma(\widehat{P}_1 - \widehat{P}_2) = \sqrt{\frac{p_1\,(1-p_1)}{n_1} + \frac{p_2\,(1-p_2)}{n_2}}$

Simulation

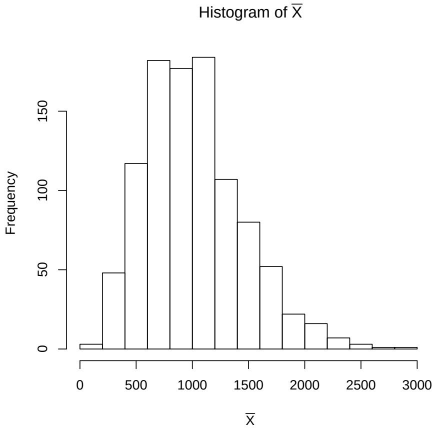
Although a proof of the central limit theorem is beyond the scope of the course, we can certainly carry out a **simulation study** in order to convince ourselves of the result.

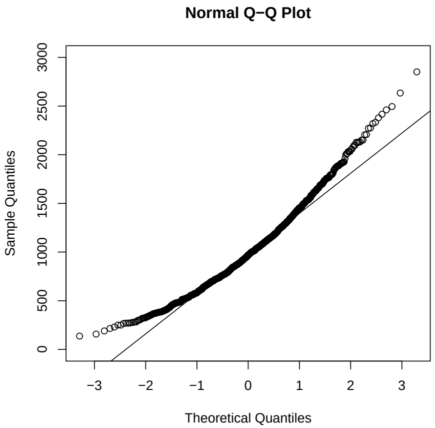
This is done as follows:

1. Generate a sample of size n from some distribution.
2. Calculate $\bar{x}$ for this sample.
3. Repeat the above two steps a number of times to create replicates of $\bar{x}$ (using a `for` loop).
4. Look at the distribution of the set of $\bar{x}$'s via a histogram (`hist`) and compare to the normal distribution using a Q-Q plot (`qqnorm`).

Example: Exponential Data

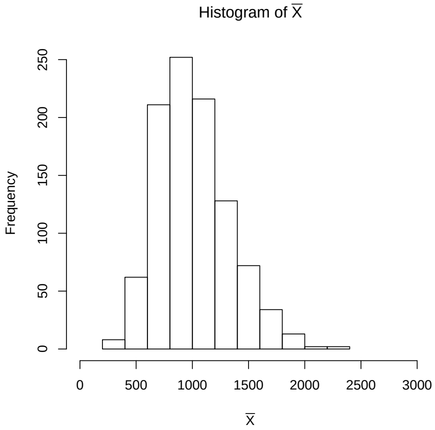
$X \sim$ Exponential$(\lambda = 0.001)$ with $\boxed{n = 5}$.

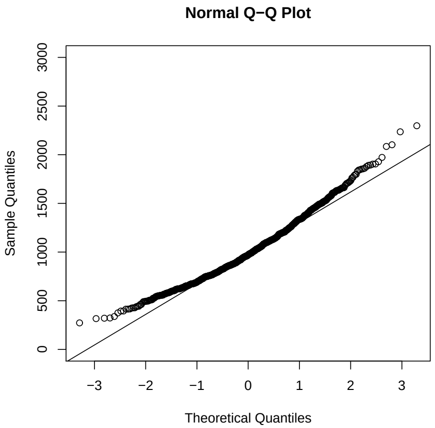




Example: Exponential Data

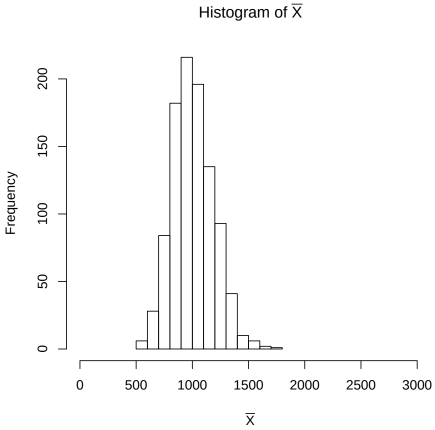
$X \sim \text{Exponential}(\lambda = 0.001)$ with $\boxed{n = 10}$.

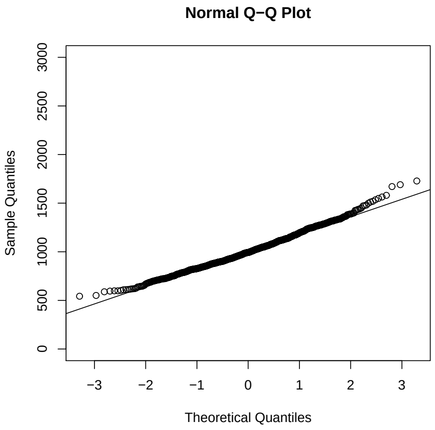




Example: Exponential Data

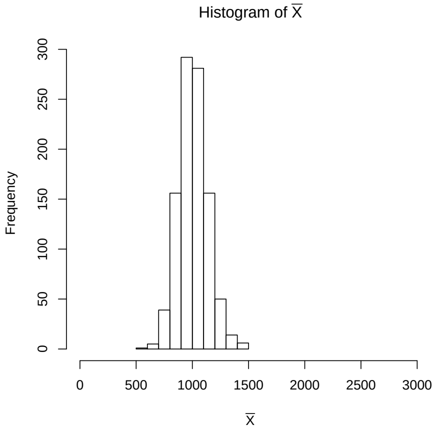
$X \sim$ Exponential$(\lambda = 0.001)$ with $\boxed{n = 30}$.

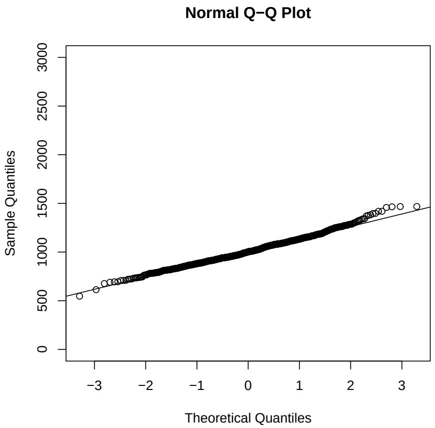




Example: Exponential Data

$X \sim \text{Exponential}(\lambda = 0.001)$ with $\boxed{n = 60}$.





R Code

Code for carrying out simulation:

```
n = 10  # sample size

set.seed(142981)

simreps = 1000   # simulation replicates
                 # (just needs to be a big number)
xbar = rep(0, simreps)

for(i in 1:simreps){
   xbar[i] = mean(rexp(n, rate=0.001))
}

hist(xbar, xlim=c(0,3000))

qqnorm(xbar, ylim=c(0,3000));qqline(xbar)
```