Statistics 301: Probability and Statistics

Sampling Distributions

Module 7

2018

Three Types of Distributions

data distribution
the distribution of a variable in a *sample*

population distribution
the *probability distribution* of a *single observation* of a variable

sampling distribution
the *probability distribution* of a *statistic*

Terms I

sampling distribution: a probability distribution of a statistic; it is a distribution of *all possible samples* (random samples) from a population and how often each outcome occurs in repeated sampling (of the same size n). Given simple random samples of size n from a given population with a measured characteristic such as mean $\overline{X}$, proportion ($\hat{p}$), or standard deviation (s) for each sample, the probability distribution of all the measured characteristics is called a sampling distribution. Use of statistic to estimate the parameter is the main function of inferential statistics as it provides the properties of the statistic.

Terms II

law of large numbers states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become ever closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order (overall), the long-term observed relative frequency will approach the theoretical probability

Central Limit Theorem (CLT)

Definition
The sampling distribution of the sample statistic is approximately normal with mean μ_X and standard deviation (of the sampling distribution of the sample mean) $se = \frac{\sigma_X}{\sqrt{n}}$, provided n is sufficiently large.

Sampling distribution of the Sample Mean

If we take n observations of a quantitative variable and then compute the mean ($\bar{x}$) of those observations in the sample, then $\bar{x}$ is the sample mean statistic.

Assumptions: Each observation x has the same probability distribution with mean μ and standard deviation σ, and the observations are *independent.*

Properties of the Sampling Distribution of $\bar{x}$

(1) The *mean* of the sampling distribution is μ

(2) The *standard deviation* of the sampling distribution is $se = \frac{\sigma}{\sqrt{n}}$

(3) The *shape* of the sampling distribution becomes more like a normal distribution as n increases

Sampling distribution of the Sample Mean

$$\overline{X} \sim N\left(\mu, se_{mean}\right)$$

$$\text{Standard error of the mean: } \sigma_{\overline{X}} = se_{mean} = \frac{\sigma}{\sqrt{n}}$$

$$z = \frac{\overline{X} - \mu}{se_{mean}}$$

Sample sizes should be $n \geq 30$ for the sample mean If a distribution is already inherently normal, the sample size stipulation can be ignored.

Sampling distribution of the Sample Proportion ($\hat{p}$)

If we make n observations, and count the number of observations on which an outcome happens (call this x), then $\hat{p} = \frac{x}{n}$ is the *sample proportion* statistic.

Assumptions: x has a binomial distribution where n is the number of trials and the probability of the outcome on each trial is p.

Properties of the Sampling Distribution of $\hat{p}$

(1) The *mean* of the sampling distribution is p.

(2) The *standard deviation* of the sampling distribution is $\sqrt{pq/n}$.

(3) The *shape* of the sampling distribution becomes more like a normal distribution as n increases.

Sampling distribution of $\hat{p}$

$$\hat{p} \sim N\left(p, se_{\hat{p}}\right)$$

$$\text{Standard error of the proportion: } \sigma_{\hat{p}} = se_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

$$z = \frac{\hat{p} - p}{se_{\hat{p}}}$$

Sample sizes should be $n \geq 60$ for the sample proportion

Sampling distribution of the Sample Sum (Total)

If we take n observations of a quantitative variable and then compute the mean total (sum) ($\hat{\tau} = n\bar{x}$) of those observations in the sample, then $\hat{\tau}$ is the sample total statistic.

Assumptions: Each observation x has the same probability distribution with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$, and the observations are *independent*.

Properties of the Sampling Distribution of τ

(1) The *mean total* of the sampling distribution is τ

(2) The *standard deviation (of the total)* of the sampling distribution is $se = \sqrt{n}\sigma$

(3) The *shape* of the sampling distribution becomes more like a normal distribution as n increases

Sampling distribution of the Sample Sum (Total)

$$\hat{\tau} = n\overline{X} \quad \tau = n\mu \quad se_{sum} = \sqrt{n}\sigma$$

$$\hat{\tau} \sim N(\tau, se_{sum})\ with\ se_{sum} = \sqrt{n}\sigma$$

$$z = \frac{n\overline{X} - n\mu}{se_{sum}} = \frac{\hat{\tau} - \tau}{se_{sum}}$$

Simulation Examples to Show CLT

To simulate the CLT and how it works, a random sample of 500 observations is taken from three different distributions: normal, exponential, and binomial. The purpose is to demonstrate the distribution of the sample mean; regardless of the original distribution, the distribution of the sample mean will be approximately normal.

CLT with normal

Simulation of a random sample of 500 observations from a normal distribution with mean of 100 and sd of 10

`rnorm()`: randomly generates values from the normal distribution in `R`

`rnorm(n,mean= ,sd= )`
- n: number of observations
- mean: mean to use for random sample
- sd: standard deviation for random sample

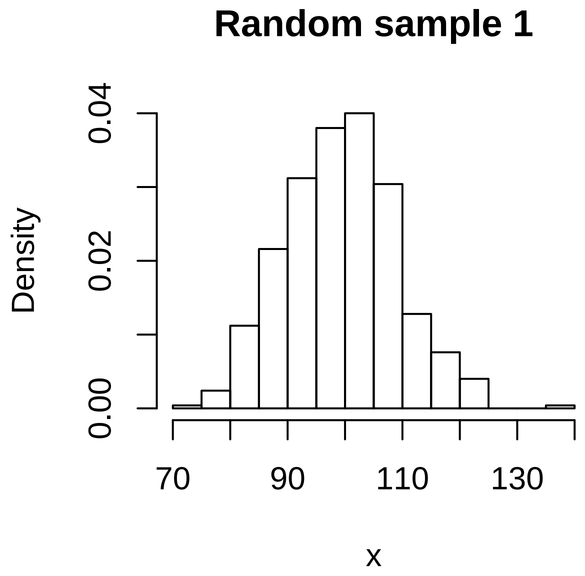
CLT with normal

Sample with $n = 500$

```
x=rnorm(500,mean=100,sd=10); cbind(mean(x),sd(x))
```

```
         [,1]     [,2]
[1,] 99.3223 9.65044
```

```r
hist(x,prob=TRUE,main='Random sample 1')
```

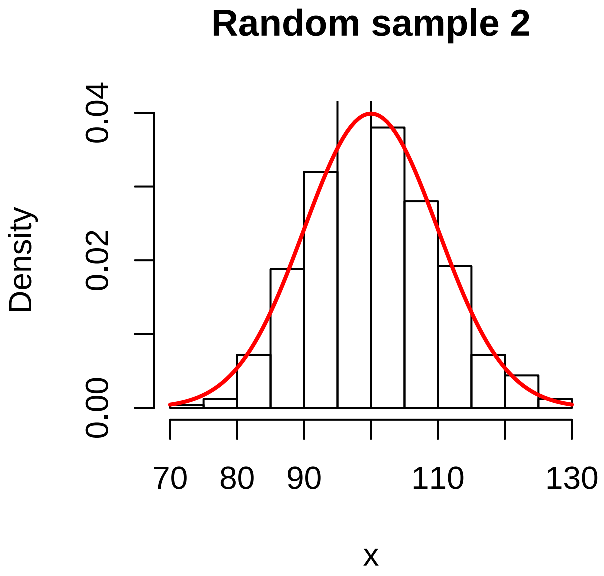


CLT with normal

```r
x=rnorm(500,mean=100,sd=10); mean(x)
```

```
[1] 100.224
```

```r
hist(x,prob=TRUE,ylim=c(0,0.04),main='Random sample 2')
curve(dnorm(x,mean=100,sd=10),70,130,add=TRUE,lwd=2,col="red")
```



CLT simulation with normal

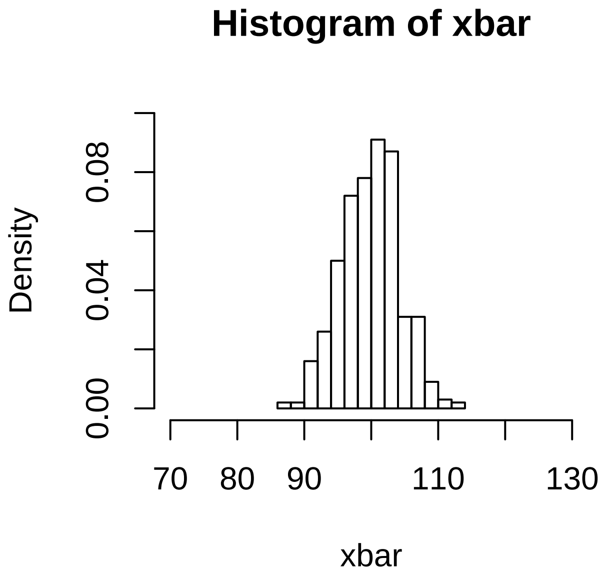
Simluation Process

- Set the mean, standard deviation and sample size

- Create empty vector to contain sample means from for-loop
- For-loop calculates the sample means from 500 simulations of sample size 5 (each sample has 5 observations and I am simulating 500 samples of size 5 and will get 500 sample means)

CLT simulation with normal

```r
mu=100; sigma=10; n=5; xbar=rep(0,500)
for (i in 1:500)
{ xbar[i]=mean(rnorm(n,mean=mu,sd=sigma)) }
hist(xbar,prob=TRUE,breaks=12,xlim=c(70,130),ylim=c(0,0.1))
```



Exponential Distribution

We will look at a random sample of 500 observations from an exponential distribution with rate of 1. The exponential distribution is one that models (describes) the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate, λ.

$$f(x) = \lambda e^{-\lambda x}$$

With

$$EX = \frac{1}{\lambda}$$

$$VX = \frac{1}{\lambda^2}$$

CLT with exponential

`rexp()`: randomly generates values from the exponential distribution in `R`

`rexp(n,rate= )`
- n: number of observations
- rate: the rate is rate=1/mean (default=1)

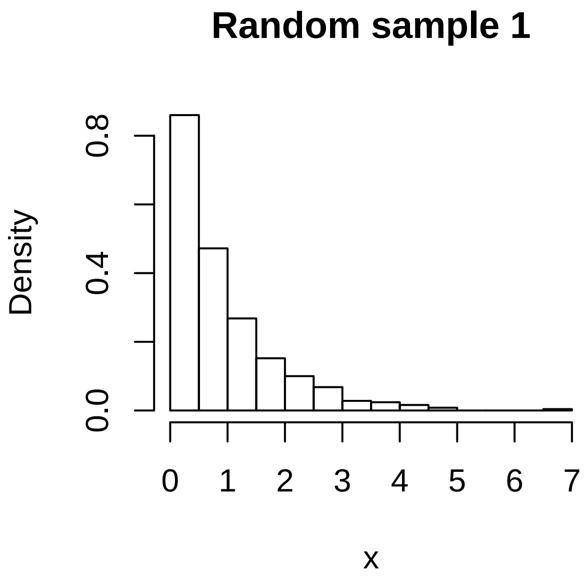
CLT with exponential

Sample with $n = 500$

```r
x=rexp(500); mean(x)
```

```
[1] 0.925837
```

```r
hist(x,prob=TRUE,main='Random sample 1')
```

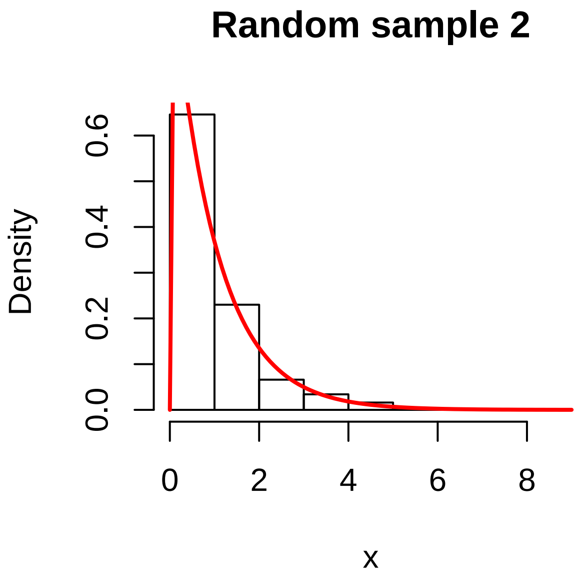


CLT with exponential

```r
x=rexp(500); mean(x)
```

```
[1] 0.970876
```

```r
hist(x,prob=TRUE,main='Random sample 2')
curve(dexp(x),add=TRUE,lwd=2,col="red")
```



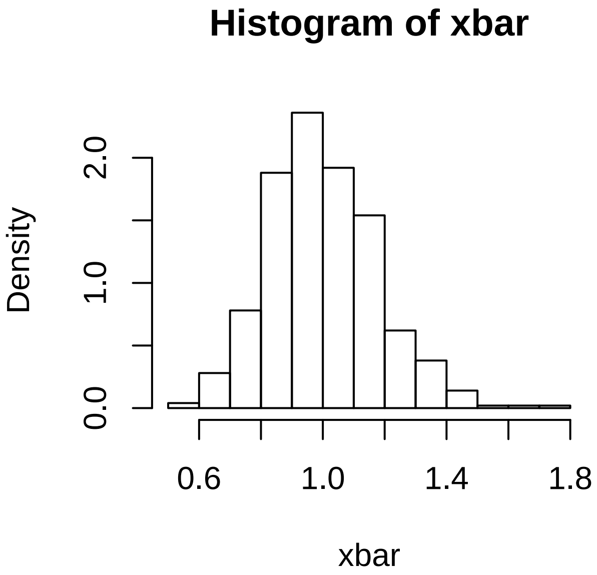
CLT simulation with exponential

Simluation Process

- Set the mean, standard deviation and sample size
- Create empty vector to contain sample means from for-loop
- For-loop calculates the sample means from 500 simulations of sample size 30 (each sample has 30 observations and I am simulating 500 samples of size 30 and will get 500 sample means)

CLT simulation with exponential

```
mu=1; sigma=1; n=30; xbar=rep(0,500)
for (i in 1:500)
{ xbar[i]=mean(rexp(n)) }
hist(xbar,prob=TRUE,breaks=12)
```



Binomial Distribution

The binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p.

We will look at a random sample of 500 observations from a binomial distribution with $p = 0.8$ $(q = 1 - p = 1 - .8 = 0.2)$ and $n = 10$

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

With

$$EX = np$$

$$VX = npq$$

CLT with binomial

`rbinom()`: randomly generates values from the binomial distribution in `R`

`rbinom(n,size= ,prob= )`
- n: number of observations
- size: number of trials
- prob: probability of success on each trial

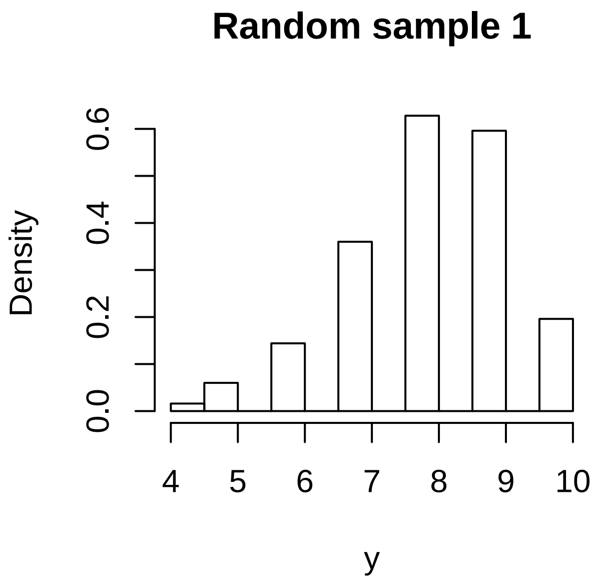
CLT with binomial

Sample with $p = 0.8$ *and size* $= 10$ $n = 500$

```
y=rbinom(500,10,.8); mean(y)
```

```
[1] 8.048
```

```
hist(y,prob=T,main='Random sample 1')
```

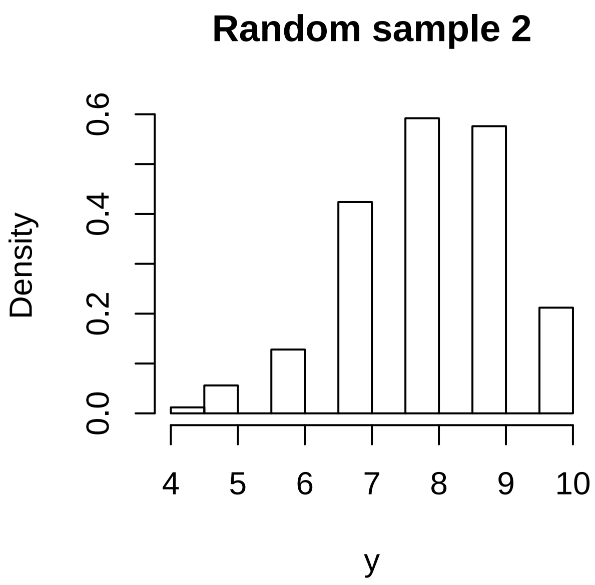


CLT with binomial

```
y=rbinom(500,10,.8); mean(y)
```

```
[1] 8.052
```

```
hist(y,prob=T,main='Random sample 2')
```



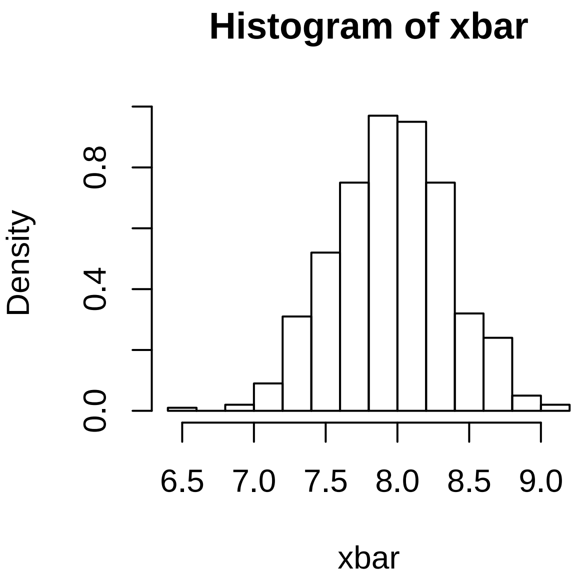
CLT simulation with binomial

Simluation Process

- Set the mean, standard deviation and sample size
- Create empty vector to contain sample means from for-loop
- For-loop calculates the sample means from 500 simulations of sample size 30 (each sample has 30 observations and I am simulating 500 samples of size 30 and will get 500 sample means)

CLT simulation with binomial

```
mu=8; sigma=1.26; n=10; xbar=rep(0,500)
for (i in 1:500)
{ xbar[i]=mean(rbinom(n,10,.8)) }
hist(xbar,prob=TRUE,breaks=15)
```



CLT for sample mean ($\overline{X}$) and sample sum/total ($\hat{\tau}$)

for sample mean ($\overline{X}$) and total ($\hat{\tau}$)

The level of a particular pollutant, nitrogen dioxide (NO_2), in the exhaust of a hypothetical model of car, that when driven in city traffic, has a mean level of 2.1 grams per mile (g/m) and a standard deviation of 0.3 g/m. Suppose a company has a fleet of 35 of these cars.

(a) What is the mean and standard deviation of the sampling distribution of the sample mean?

mean: $\mu_X = \mu = 2.1$ and $se_{mean} = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{\sqrt{35}} = 0.0507$

$\overline{X} \sim N(\mu, se_{mean}) = \overline{X} \sim N(2.1, 0.0507)$

CLT for $\overline{X}$ and $\hat{\tau}$ solutions

(b) find the probability that the mean NO_2 level is less than 2.03 g/m

$$P(\overline{X} < 2.03) = P\left(Z < \frac{2.03 - 2.1}{0.0507}\right) = P(Z < -1.38) = 0.083793$$

(c) Mandates by the EPA state that the average of the fleet of these cars cannot exceed 2.2 g/m, find the probability that the fleet NO_2 levels from their fleet exceed the EPA mandate

$$P(\overline{X} > 2.2) = 1 - P\left(Z < \frac{2.2 - 2.1}{0.0507}\right)$$
$$= 1 - P(Z < 1.97) = 1 - 0.975581 = 0.024419$$

CLT for $\overline{X}$ and $\hat{\tau}$ solutions

(d) At most, 25% of these cars exceed what *mean* NO_2 value?

Find the z score that represents the top 25%, which is the same as the bottom 75% (is also $Q3$) and what is needed to find $z_{0.75} = 0.67449$. Next use $z = \frac{\overline{X}-\mu}{se_{mean}}$ and solve for $\overline{X}$: $\overline{X} = z(se_{mean}) + \mu$

$$\overline{X} = (0.67449)(0.0507) + 2.1 = 2.134197$$

CLT for $\overline{X}$ and $\hat{\tau}$ solutions

(e) what is the mean and standard deviation of the total amount (sum), in g/m, of NO_2 in the exhaust for the fleet?

$$\tau = n\mu = 35(2.1) = 73.5$$

$$se_{sum} = \sqrt{n}\sigma = \sqrt{35}(0.3) = 1.774824$$

$$\hat{\tau} \sim N(\tau, se_{sum}) = \hat{\tau} \sim N(73.5, 1.7748)$$

CLT for $\overline{X}$ and $\hat{\tau}$ solutions

(f) find the probability that the total amount of NO_2 for the fleet is between 70 and 75 g/m

$$P(70 < \hat{\tau} < 75) = P\left(\frac{70 - 73.5}{1.7448} < Z < \frac{75 - 73.5}{1.7748}\right)$$
$$= P(-2.01 < Z < 0.86) = P(Z < 0.86) - P(Z < -2.01)$$
$$= 0.805105 - 0.022216 = 0.78289$$

CLT for proportion ($\hat{p}$)

Mars company claims that 10% of the M&M's it produces are green. Suppose that candies are packaged at random in bags that contain 60 candies.
(a) Describe the sampling distribution of the sample proportion (what should the distribution look like?); calculate the mean proportion and standard deviation of the sampling distribution of the sample proportion of green M&M's in bags that contain 60 candies (calculate p and se).
(b) What is the probability that a bag of 60 candies will have more than 13% green M&M's?

CLT for $\hat{p}$ solutions

(a) Describe the sampling distribution of the sample proportion; calculate the mean proportion and standard deviation of the sampling distribution of the sample proportion of green M&M's in bags that contain 60 candies.

The distribution of the sample proportion will be approximately normal since $n \geq 60$. The mean proportion $p = 0.1$ and the standard error is $\sqrt{\frac{pq}{n}} = \sqrt{\frac{(.1)(.9)}{60}} = 0.0387$ (the standard deviation of the sampling distribution of the sample proportion). Thus

$$\hat{p} \sim N(0.1, 0.0387)$$

CLT for $\hat{p}$ solutions

(b) What is the probability that a bag of 60 candies will have more than 13% green M&M's?

$$P(\hat{p} > 0.13) = P\left(Z > \frac{0.13 - 0.1}{0.0387}\right)$$
$$= P(Z > 0.78) = 1 - P(Z < 0.78) = 1 - 0.782305$$
$$= 0.2177$$