Math 1710 Class 17

Sampling Distributions, CLT

Dr. Back

Oct. 5, 2009

Sampling Distribution of a Statistic

Given n observations $x_1, x_2, \ldots, x_n$, a statistic is any number determined by these observations.

Sampling Distribution of a Statistic

Given n observations $x_1, x_2, \ldots, x_n$, a statistic is any number determined by these observations.

For example the mean $\bar{x}$ or standard deviation s.

Sampling Distribution of a Statistic

"Sampling Distribution of a Statistic"

A probability model describing the chance of different values of the statistic showing up.

Three CLT problems

(1) 80% of all cars on the interstate speed.
Randomly sample 50.
What proportion of speeders might we see?

Three CLT problems

(1) 80% of all cars on the interstate speed.
Randomly sample 50.
What proportion of speeders might we see?
(e.g.: A policeman has a quota of at least 35 speeders to ticket.
What is the chance he'll meet his quota in such a sample?)

Three CLT problems

(2) At birth, babies average 7.8 pounds with a standard deviation of 2.1 pounds.
34 Babies born near a large possibly polluting factory average 7.2 pounds.

Three CLT problems

(2) At birth, babies average 7.8 pounds with a standard deviation of 2.1 pounds.
34 Babies born near a large possibly polluting factory average 7.2 pounds.
Is that unusually low?

Three CLT problems

(3) Suppose SAT's have a mean of 500, standard deviation of 100 and are approximately normal.
Form means of samples of 20 students.

Three CLT problems

(3) Suppose SAT's have a mean of 500, standard deviation of 100 and are approximately normal.
Form means of samples of 20 students.
What distribution do these follow? What is the chance of a sample averaging over 600? Over 550?

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.
$X \sim \text{Binomial(n,p)} = \text{Binomial}(100, .7)$

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.
$X \sim$ Binomial(n,p) $=$ Binomial(100, .7)
$\mu = np = 70$, $\sigma = \sqrt{npq} = \sqrt{21} = 4.58$.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.
$X \sim$ Binomial(n,p) = Binomial(100, .7)
$\mu = np = 70$, $\sigma = \sqrt{npq} = \sqrt{21} = 4.58$.
X approximated by normal RV $Y \sim N(70, 4.58)$.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.
X approximated by normal RV $Y \sim N(70, 4.58)$.
Now suppose we keep track of the observed proportion

$$\hat{p} = \frac{X}{n} = \frac{X}{100}$$

here.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.
X approximated by normal RV $Y \sim N(70, 4.58)$.
Now suppose we keep track of the observed proportion

$$\hat{p} = \frac{X}{n} = \frac{X}{100}$$

here.
Then $\hat{p}$ is an RV approximated by
$Y/n \sim N(p, \sqrt{\frac{pq}{n}}) = N(.7, .0458)$.

Proportion Case of the CLT

Suppose true proportion approving the president is p=70%.
Poll n=100 people.
Let the RV X describes the number who approve.
X approximated by normal RV $Y \sim N(70, 4.58)$.
Now suppose we keep track of the observed proportion

$$\hat{p} = \frac{X}{n} = \frac{X}{100}$$

here.
Then $\hat{p}$ is an RV approximated by
$Y/n \sim N(p, \sqrt{\frac{pq}{n}}) = N(.7, .0458)$.
Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.
Your textbook uses the notation

$$\text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.
Your textbook uses the notation

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

Note that $\hat{p}$ is on the left side but not the right!
This is NOT a functional notation.

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.
Your textbook uses the notation

$$\text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

$\text{SD}(\hat{p})$ stands for Standard Deviation of the Sampling Distribution of $\hat{p}$.

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

$\mathrm{SD}(\hat{p})$ stands for Standard Deviation of the Sampling Distribution of $\hat{p}$.
Your textbook also uses the notation

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

$\mathrm{SD}(\hat{p})$ stands for Standard Deviation of the Sampling Distribution of $\hat{p}$.

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

$\mathrm{SE}(\hat{p})$ stands for the Standard Error of $\hat{p}$.

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.

$$\text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

$\text{SD}(\hat{p})$ stands for Standard Deviation of the Sampling Distribution of $\hat{p}$.

$$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

$\text{SE}(\hat{p})$ stands for the Standard Error of $\hat{p}$.
SE is just an approximate version of SD.

Proportion Case of the CLT

Thus $N(p, \sqrt{\frac{pq}{n}})$ is the sampling distribution of $\hat{p}$.

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}.$$

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

$\mathrm{SE}(\hat{p})$ stands for the Standard Error of $\hat{p}$.
SE is just an approximate version of SD.
Many people prefer to use the word *Standard Error* for both SD and SE.
SD is exact. SE is approximate.
The idea of such people is to use the term Standard Error for any standard deviation of a sampling distribution.
But the authors of our textbook do not go this way.

A Proportion CLT problems

(1) 80% of all cars on the interstate speed.
Randomly sample 50.
What proportion of speeders might we see?

A Proportion CLT problems

(1) 80% of all cars on the interstate speed.
Randomly sample 50.
What proportion of speeders might we see?
(e.g.: A policeman has a quota of at least 35 speeders to ticket.
What is the chance he'll meet his quota in such a sample?)

A Proportion CLT problems

(1) 80% of all cars on the interstate speed.
Randomly sample 50.
What proportion of speeders might we see?
(e.g.: A policeman has a quota of at least 35 speeders to ticket.
What is the chance he'll meet his quota in such a sample?)
Solution: Let p be the true proportion of speeders.

A Proportion CLT problems

(1) 80% of all cars on the interstate speed.
Randomly sample 50.
What proportion of speeders might we see?
(e.g.: A policeman has a quota of at least 35 speeders to ticket.
What is the chance he'll meet his quota in such a sample?)
Solution: Let p be the true proportion of speeders.
We are given $p = .8$. So

$$SD(\hat{p}) = \sqrt{\frac{.8 \cdot .2}{50}} = .057$$

A Proportion CLT problems

(e.g.: A policeman has a quota of at least 35 speeders to ticket. What is the chance he'll meet his quota in such a sample?)

Solution: Let p be the true proportion of speeders.

We are given $p = .8$. So

$$SD(\hat{p}) = \sqrt{\frac{.8 \cdot .2}{50}} = .057$$

And the sampling distribution of $\hat{p}$ is $N(.8, .057)$.

A Proportion CLT problems

(e.g.: A policeman has a quota of at least 35 speeders to ticket. What is the chance he'll meet his quota in such a sample?)

Solution: Let p be the true proportion of speeders.
We are given $p = .8$. So

$$SD(\hat{p}) = \sqrt{\frac{.8 \cdot .2}{50}} = .057$$

And the sampling distribution of $\hat{p}$ is $N(.8, .057)$.
The z score of $\frac{35}{50} = .7$ is

$$\frac{-.1}{.057} = -1.75$$

A Proportion CLT problems

(e.g.: A policeman has a quota of at least 35 speeders to ticket. What is the chance he'll meet his quota in such a sample?)

Solution: Let p be the true proportion of speeders.
We are given $p = .8$. So

$$SD(\hat{p}) = \sqrt{\frac{.8 \cdot .2}{50}} = .057$$

And the sampling distribution of $\hat{p}$ is $N(.8, .057)$.
The z score of $\frac{35}{50} = .7$ is

$$\frac{-.1}{.057} = -1.75$$

So $P(\hat{p} \geq 35) = P(Z > -1.75) = .9599$.
Assuming independence here might not be reasonable.
If one person speeds, her neighbors might be more likely to as well.

Quantitative Case of the CLT

- Suppose X is a random variable with mean μ and standard deviation σ.
- Let X_1, X_2,...,X_n be n independent copies of X.
- Set

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

- Then
  1. The mean of $\bar{X}$ is μ.
  2. The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.
  3. $\bar{X}$ is approximately normal as n gets large.

Quantitative CLT - Why?

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

1. The mean of $\bar{X}$ is μ.
2. The standard deviation of $\bar{X}$ is $\dfrac{\sigma}{\sqrt{n}}$.
3. $\bar{X}$ is approximately normal as n gets large.

Quantitative CLT - Why?

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

1. The mean of $\bar{X}$ is μ.
2. The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.
3. $\bar{X}$ is approximately normal as n gets large.

(1) is immediate from

$$E(X_1 + X_2 + \ldots + X_n) = nE(X)$$

so

$$E(\bar{X}) = \frac{nE(X)}{n} = E(X).$$

Quantitative CLT - Why?

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

1. The mean of $\bar{X}$ is μ.
2. The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.
3. $\bar{X}$ is approximately normal as n gets large.

(2) is immediate from

$$Var(X_1 + X_2 + \ldots + X_n) = nVar(X)$$

so

$$Var(\bar{X}) = \frac{nVar(X)}{n^2} = \frac{Var(X)}{n}.$$

And so the std. dev of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.

Quantitative CLT - Why?

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

1. The mean of $\bar{X}$ is μ.
2. The standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.
3. $\bar{X}$ is approximately normal as n gets large.

(3) is much deeper! We will illustrate, but it's only barely possible to prove in even an advanced undergraduate course. Your text uses the notation

$$SD(\bar{x}) = \frac{\sigma}{n}$$

for the exact standard deviation of the sampling distribution of $\bar{x}$ and $SE(\bar{x}) = \frac{s}{n}$ for its approximation, the standard error of $\bar{x}$.