University of Washington

STATISTICS

Elements of Statistical Methods
Lots of Data or Large Samples (Ch 8)

Fritz Scholz

Spring Quarter 2010

February 26, 2010

$\bar{x}_n$ and $\bar{X}_n$

We introduced the sample mean $\bar{x}_n$ as the average of the observed sample values $\vec{x} = \{x_1, \ldots, x_n\}$, using the plug-in principle.

In parallel, we can also consider the average of the random variables $X_1, \ldots, X_n$

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and view it as a random variable $\bar{X}_n : S \longrightarrow R$.

$\bar{X}_n$ is just the average of n such functions $X_i : S \longrightarrow R$.

Since the x_i are just the observed values of the X_i, we can view

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{as the observed value of} \quad \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$\bar{X}_n$, viewed as a random variable, has a distribution. Let us experiment!

R makes it very easy to get hands on experience.

Behavior of $\bar{X}_n$ when Sampling $\chi^2(3)$

Take a sample of size $n = 5$ from $\chi^2(3)$ via `x <- rchisq(5,3)`

and then compute its mean `mean(x)`. $X \sim \chi^2(3) \Rightarrow \mu = EX = 3$.

Repeat this several times, i.e., get several observed values $\bar{x}_5$ of $\bar{X}_n$.

```
> x <- rchisq(5,3)
> mean(x)
[1] 2.199718
> x <- rchisq(5,3)
> mean(x)
[1] 4.647138
> x <- rchisq(5,3)
> mean(x)
[1] 4.771858
```

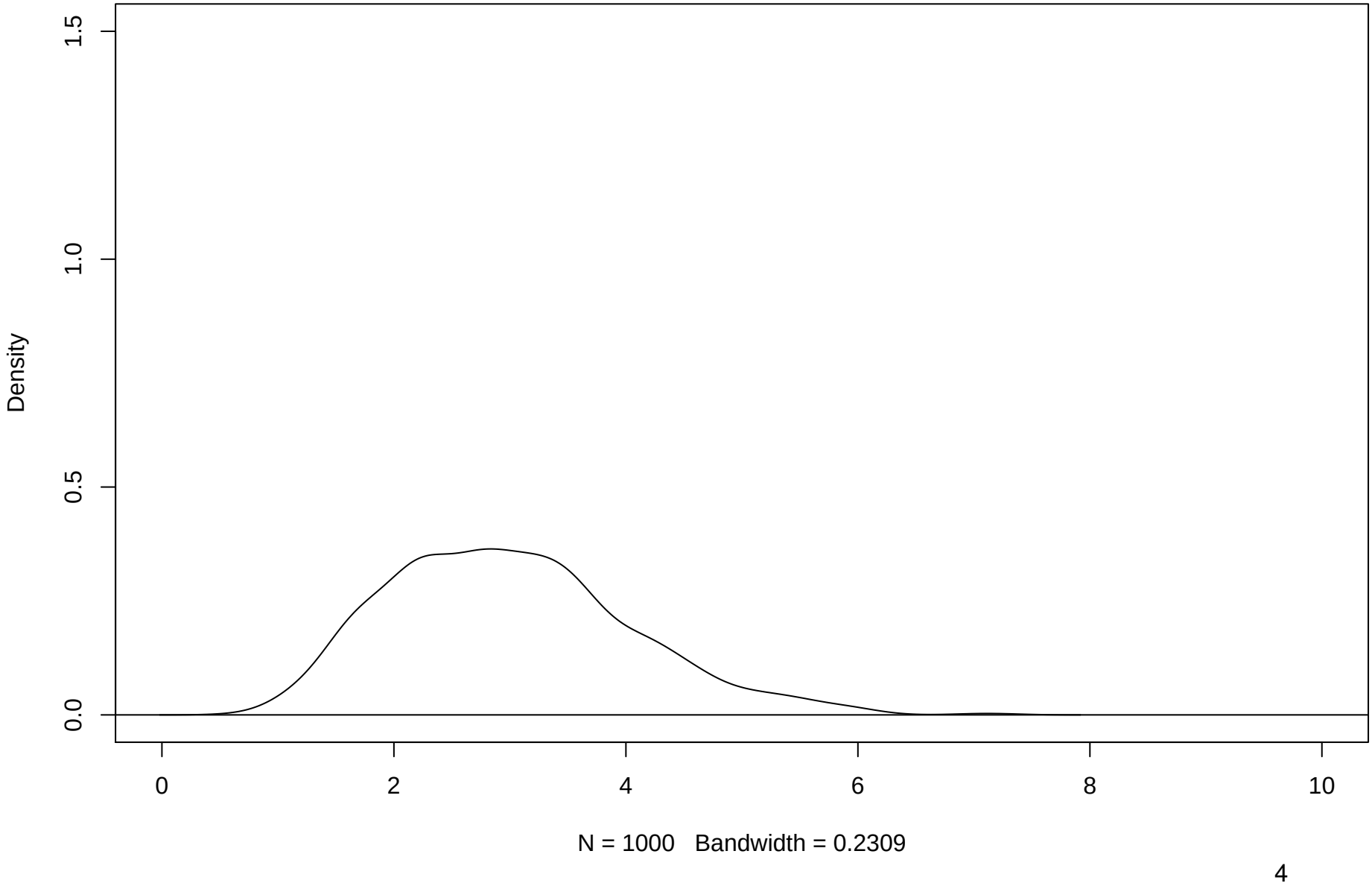
These values scatter widely around $\mu = 3$, sampling variability!

Sampling Variability of $\bar{X}_5$

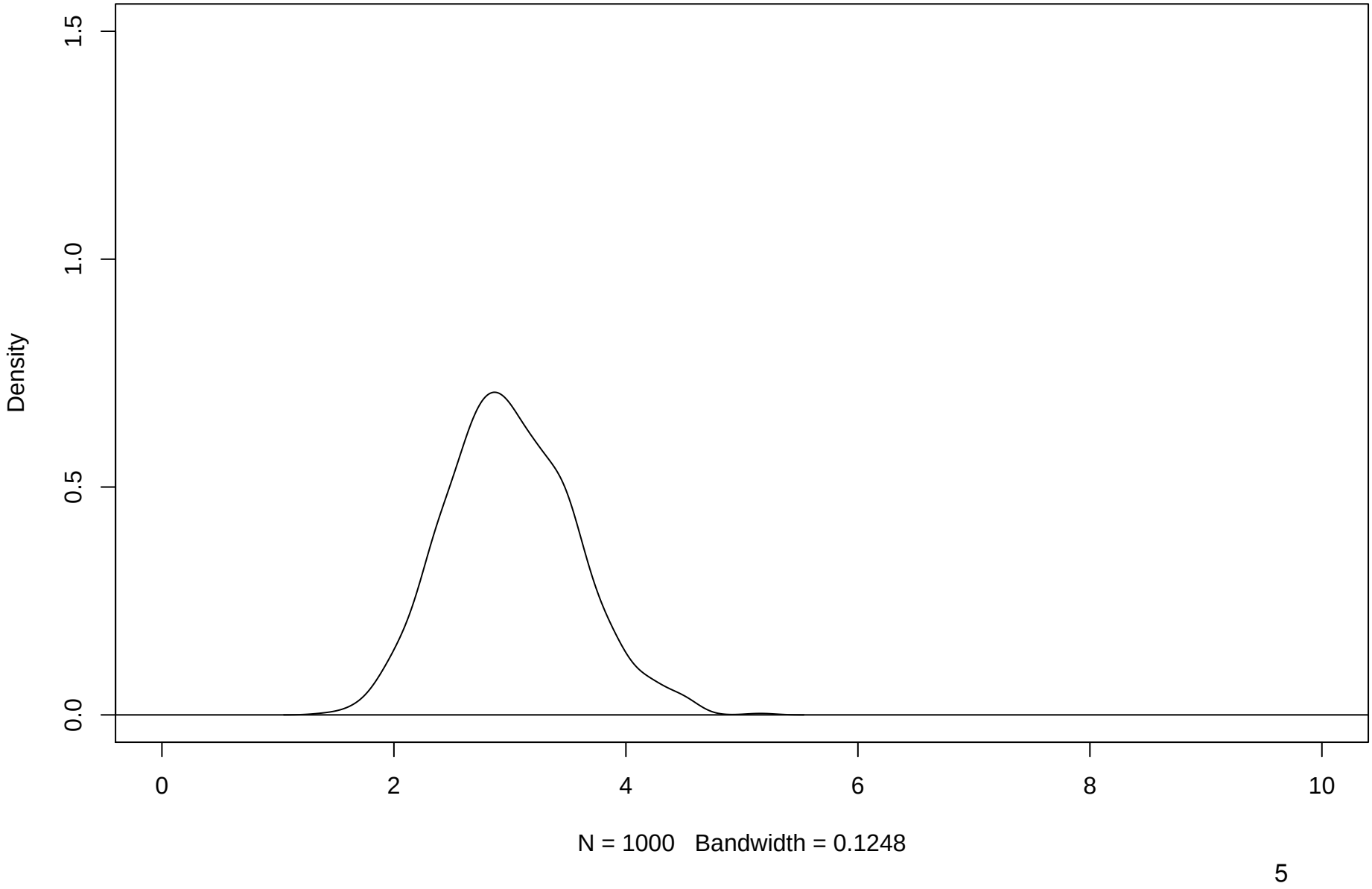
To get a less haphazard view of this sampling variability, we repeat this process $N_{\text{sim}} = 1000$ times and look at these 1000 observed sample means using a kernel density plot. This is best implemented in a `function` with a loop.

```
chi2averageSim <- function(Nsim=1000,n=5,k=3){
Xbar <- numeric(Nsim)
for( i in 1:Nsim ){
    x <- rchisq(n,k); Xbar[i] <- mean(x)
}
plot(density(Xbar),xlim=c(0,9),ylim=c(0,1.4),main="")
abline(h=0)
}
```

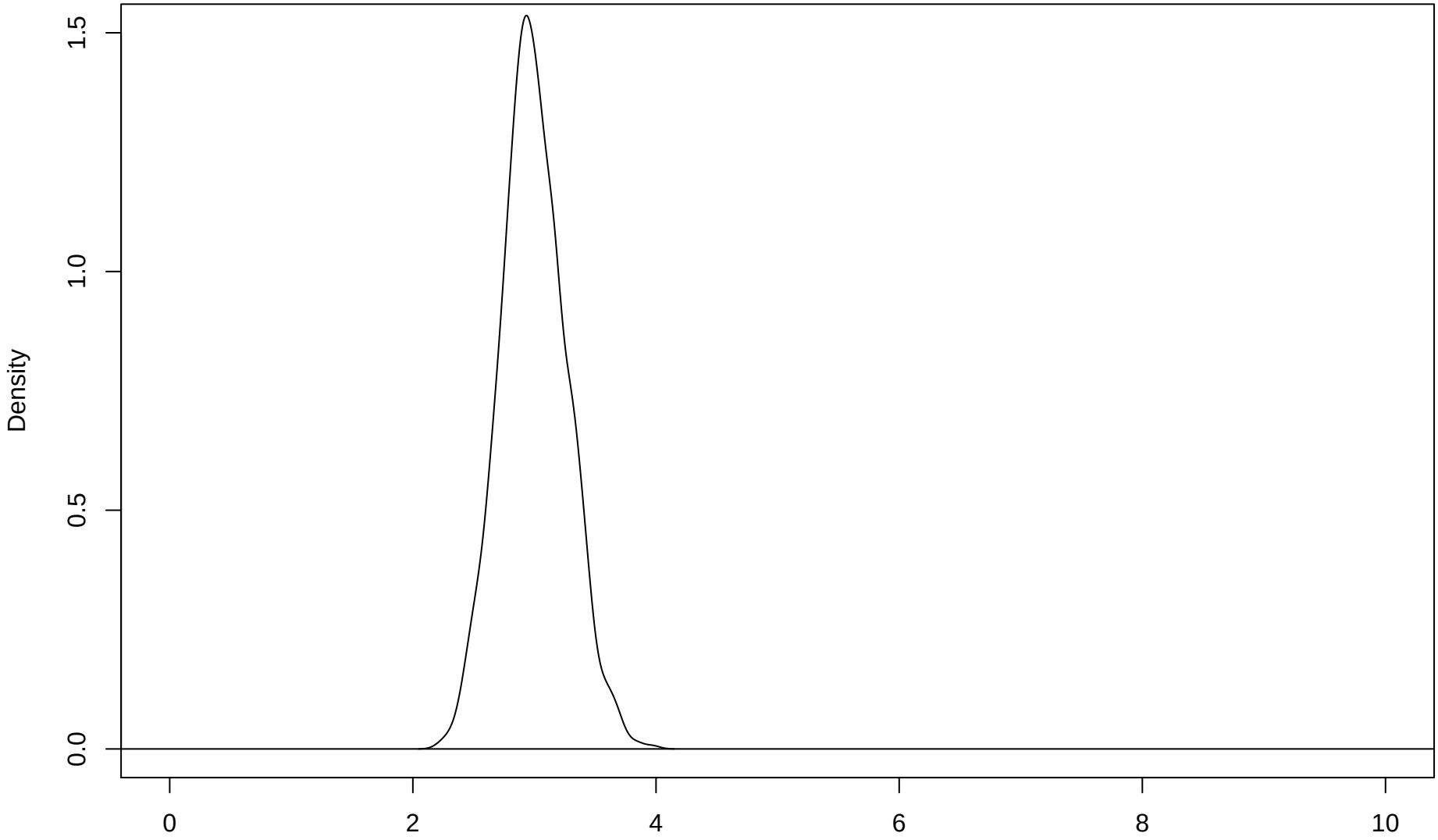
Sampling Variation of $\bar{X}_5$: $X_i \sim \chi^2(3)$



Sampling Variation of $\bar{X}_{20}$: $X_i \sim \chi^2(3)$



Sampling Variation of $\bar{X}_{80}$: $X_i \sim \chi^2(3)$



Comments

All three kernel density plots are on the same horizontal and vertical scale.

We see that they are all centered more or less on $\mu = 3$.

The sampling variability of $\bar{X}_n$, as we go from $n = 5$ to $n = 20$ to $n = 80$, decreases visibly, almost by a factor of $2 = \sqrt{4}$ each time (for a reason).

The mild skew to the right for $n = 5$ seems to disappear as n gets larger.

The distributions start to look more normal for larger n.

Experiment with `chi2averageSim(Nsim=1000,n=5,k=3)`, replacing $n = 5$ by $n = 20$ and $n = 80$.

Averaging Decreases Variation in $\bar{X}_n$ Distribution

What we saw experimentally, when sampling from $\chi^2(3)$, we will now generalize.

Let $X_, \ldots, X_n$ be i.i.d. $\sim F$, some cdf

with finite mean $\mu = EX_i$ and finite variance $\sigma^2 = \operatorname{var} X_i$.

$$E\bar{X}_n = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\cdot n \cdot \mu = \mu \quad \text{(independence not used)}$$

i.e., the mean of the $\bar{X}_n$ population is the same as that of the sampled population.

$$\operatorname{var}\bar{X}_n = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var} X_i = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n} \quad \text{(independence is used)}$$

$\sigma(\bar{X}_n) = \sigma/\sqrt{n}$, i.e., quadrupling n cuts $\sigma(\bar{X}_n)$ by a factor 2: $1/\sqrt{4n} = 1/(2\sqrt{n})$.

The Weak Law of Large Numbers (WLLN)

Recall our previous definition of convergence: $y_n \longrightarrow c$ as $n \longrightarrow \infty$ iff

for any $\varepsilon > 0$ we can find a natural number N such that

$$y_n \in (c-\varepsilon, c+\varepsilon) \quad \text{for all } n \geq N$$

We now replace the number sequence y_n by a sequence Y_n of random variables.

Definition: A sequence of random variables $\{Y_n\}$ converges in probability to a constant c, written $Y_n \xrightarrow{P} c$, iff for any $\varepsilon > 0$,

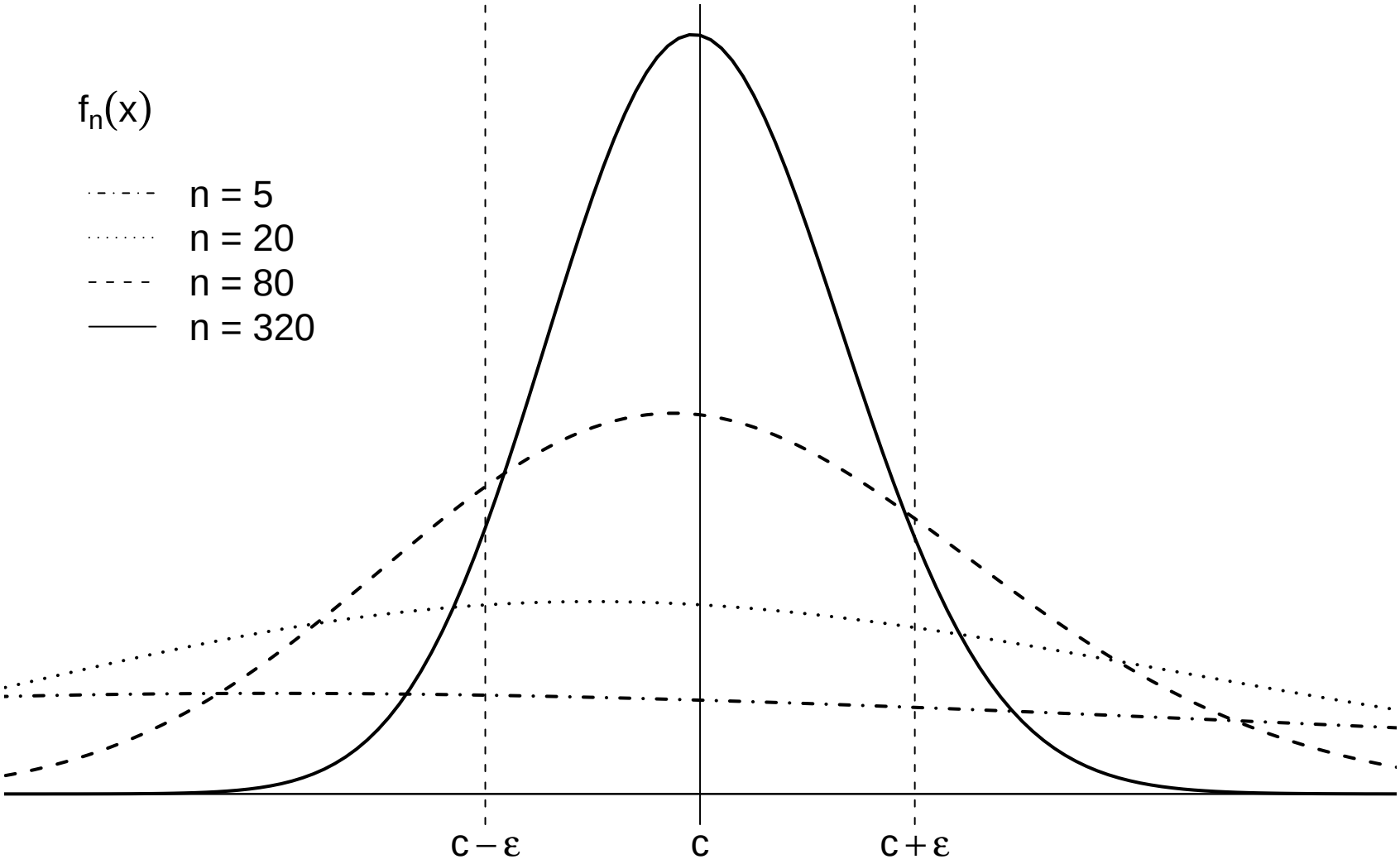
$$\lim_{n\to\infty} P(Y_n \in (c-\varepsilon, c+\varepsilon)) = 1$$

i.e., Y_n gets arbitrarily close to c with probability closer and closer to 1 as $n \longrightarrow \infty$.

In the continuous case, denoting the density of Y_n by f_n

$$\int_{c-\varepsilon}^{c+\varepsilon} f_n(x)dx = \text{Area}_{(c-\varepsilon,c+\varepsilon)}(f_n) \longrightarrow 1 \quad \text{as} \quad n \longrightarrow \infty$$

Density $f_n(x)$ of $\bar{X}_n$



The Weak Law of Large Numbers (WLLN)

Theorem (WLLN): Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables with finite mean μ and finite variance σ^2. Then

$$\bar{X}_n \xrightarrow{P} \mu \qquad \text{or equivalently} \qquad \bar{X}_n - \mu \xrightarrow{P} 0 \qquad \text{as} \qquad n \longrightarrow \infty$$

The average $\bar{X}_n$ of more and more observations X_i will get closer and closer to the mean μ of the sampled population with probability tending to 1.

Large sample sizes are good!

That is why it is important to report the sample size used in surveys.

The Frequentist's Basis for Interpretation of Probability

Corollary: Let A be an event and consider a sequence of independent and identical experiments for which we record whether the event A occurs or not.

Let $p = P(A)$ and define i.i.d. Bernoulli random variables

$$X_i = \begin{cases} 1 & A \text{ occurs} \\ 0 & A^c \text{ occurs} \end{cases}$$

Then $\bar{X}_n$ is the relative frequency with which the event A occurs in n trials.

Since $\mu = EX_i = E\bar{X}_n = p = P(A)$

$$\text{WLLN} \quad \Longrightarrow \quad \bar{X}_n \xrightarrow{P} p \quad \text{as } n \longrightarrow \infty.$$

Thus the axiomatic model of probability enriched by the concept of independence proves the frequentist's interpretation of probability.

Empirical Probabilities and Plug-In Principle

Recall that we defined the empirical probability of observing a random variable X_i with value x_i in event $A \subset R$ as

$$\hat{P}_n(A) = \frac{\#\{x_i \in A\}}{n} \qquad (\ = \hat{p}_n(A) \quad \text{may be more appropriate notation.})$$

When viewing this in terms of X_i instead of x_i we have

$$\hat{P}_n(A) = \frac{\#\{X_i \in A\}}{n} \xrightarrow{P} p = P(A) \quad \text{as } n \longrightarrow \infty.$$

The WLLN gives us a justification for approximating $P(A)$ by $\hat{P}_n(A)$.

This is often referred to as the fundamental theorem of statistics, especially when using $A = (-\infty, a]$ and then

$$\hat{F}_n(a) = \frac{\#\{X_i \leq a\}}{n} \xrightarrow{P} F(a) = P(X_i \leq a) \quad \text{as} \quad n \longrightarrow \infty.$$

Standardization of a Random Variable

A random variable X with finite mean $\mu = EX$ and finite variance σ^2 is in its standardized form Z when $Z = (X - \mu)/\sigma$

$$\begin{aligned} EZ &= \frac{E(X-\mu)}{\sigma} = \frac{\mu - \mu}{\sigma} = 0 \\ \operatorname{var} Z &= \frac{1}{\sigma^2}\operatorname{var}(X-\mu) = \frac{1}{\sigma^2}\operatorname{var}(X) = \frac{\sigma^2}{\sigma^2} = 1 \end{aligned}$$

The following are the standardized versions of X_i, $X_1 + \ldots + X_n$ and $\bar{X}_n$

random variable	expected value	standard deviation	standard units
X_i	μ	σ	$X_i - \mu)/\sigma$
$X_1 + \ldots + X_n$	$n\mu$	$\sqrt{n}\sigma$	$(\sum_{i=1}^n X_i - n\mu)/(\sqrt{n}\sigma)$
$\bar{X}_n$	μ	$\sigma/\sqrt{n}$	$(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$

Note the equivalence $(\bar{X}_n - \mu)/(\sigma/\sqrt{n}) = (\sum_{i=1}^n X_i - n\mu)/(\sqrt{n}\sigma)$

Comments on Standardization

The basic shape of the distribution remains unchanged by standardization.

$X \sim$ Bernoulli$(0.5) \Rightarrow \mu = p = 0.5$ and $\sigma = \sqrt{p(1-p)} = 0.5$,
then $Z = (X - 0.5)/0.5 = 2X - 1$ takes on the two values
$(1 - 0.5)/0.5 = 1$ and $(0 - 0.5)/0.5 = -1$ with equal probability $p = 0.5$.

Standardization $\not\Rightarrow$ the standardized random variable is normally distributed.

This misconception may come from the frequent interchangeable language usage of standardization and normalization (normal in the sense of normative).

Standardization only turns $X \sim \mathcal{N}(\mu, \sigma^2)$ into a $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$
i.e., you start with normality and you end up with normality.

Approximate distributional normality is due to different effects.

$\bar{X}_n - \mu \xrightarrow{P} 0$ (Speed?)

We have

$$\text{var}\left[\sqrt{n}\cdot(\bar{X}_n-\mu)\right] = n\cdot\text{var}(\bar{X}_n-\mu) = n\cdot\text{var}\,\bar{X}_n = n\cdot\frac{\sigma^2}{n} = \sigma^2$$

$\bar{X}_n - \mu$, multiplied by the factor $a_n = \sqrt{n}$, has mean zero and fixed variance σ^2.

Thus it appears that $a_n = \sqrt{n}$ is just the right factor to counteract the collapse,

i.e., we can view $1/\sqrt{n}$ as the rate of the collapse of $\bar{X}_n - \mu$ to zero.

Aside from a stable mean zero and variance σ^2 for $\sqrt{n}\cdot(\bar{X}_n-\mu)$,

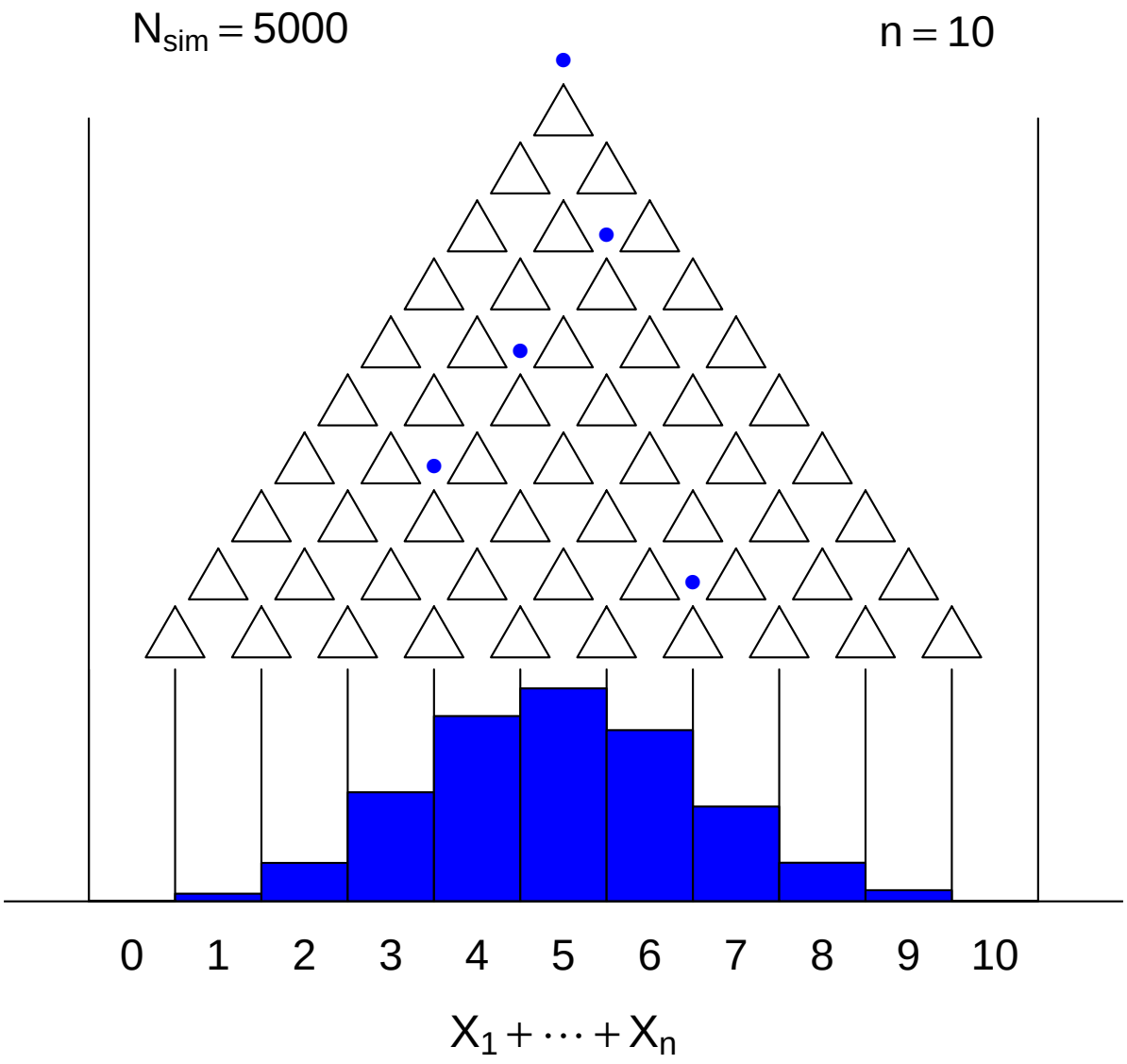
can we say more about its distribution as $n \longrightarrow \infty$?

This question is addressed by the Central Limit Theorem (CLT).

A mechanical display of the CLT, the Galton Board or quincunx,

is on display at the Pacific Science Center.

Galton Board



The Central Limit Theorem

Theorem: Let $X_1, \ldots, X_n$ i.i.d. $\sim F$, with finite mean μ and finite variance σ^2. Denote the cdf of the standardized random variables $\bar{X}_n$ and $X_1 + \ldots + X_n$, i.e.,

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + \ldots + X_n - n\mu}{\sqrt{n}\,\sigma}, \qquad \text{by} \quad F_n$$

Then for all $z \in R$

$$P(Z_n \leq z) = F_n(z) \longrightarrow \Phi(z) \qquad \text{as} \quad n \longrightarrow \infty$$

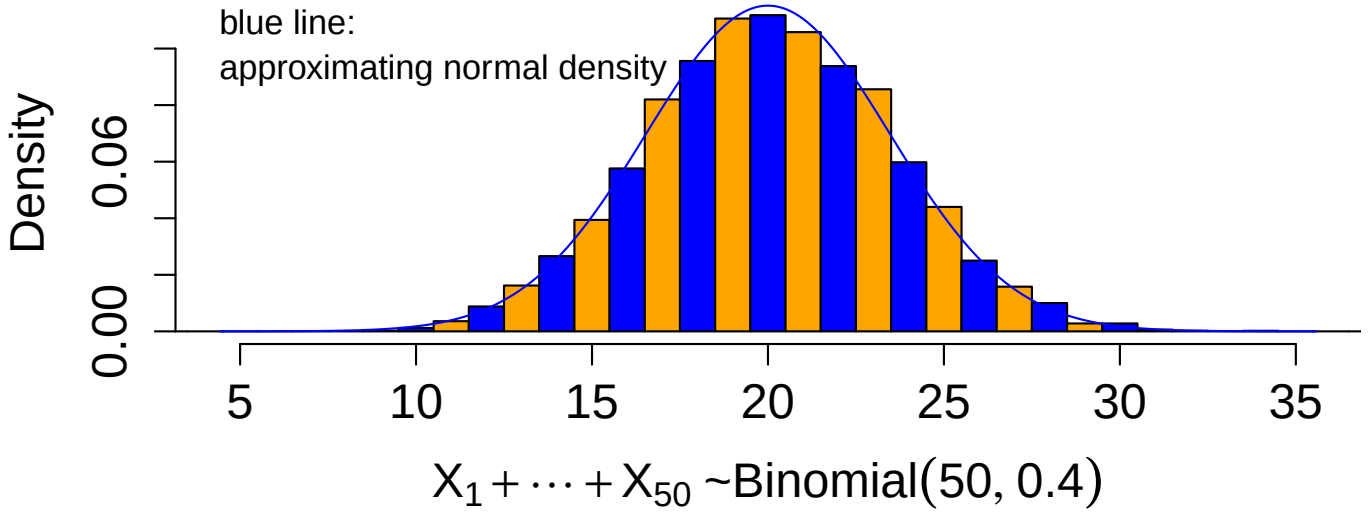
The distribution F of the X_i can be any distribution with finite μ and σ^2.

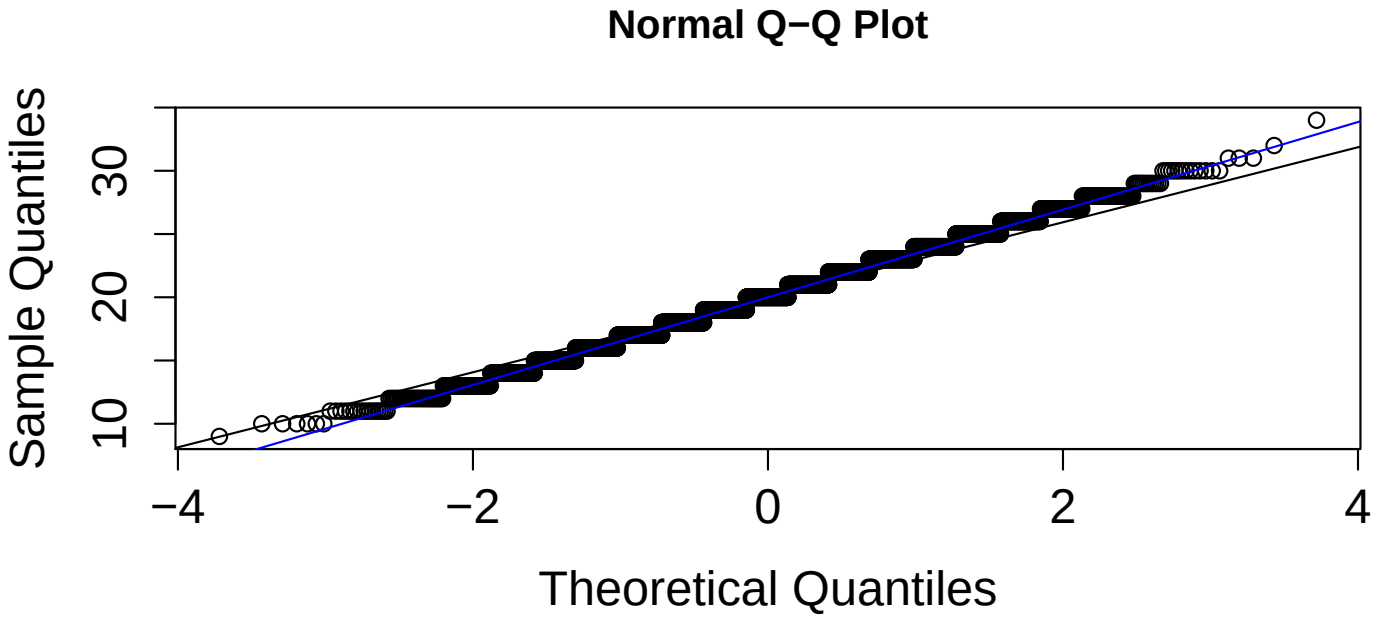
We also write $F_n(z) \approx \Phi(z)$ or $Z_n \approx \mathcal{N}(0,1)$ to express this approximation result.

Note that

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \approx \mathcal{N}(0,1) \quad \text{or} \quad \bar{X} - \mu \approx \mathcal{N}(0, \sigma^2/n) \text{ the collapse}$$

CLT for Binomial = Sum of Bernoulli R.V.s





Speed of Convergence in CLT

How fast is the convergence $F_n(z) \longrightarrow \Phi(z)$ in relation to n?

Under mild conditions: $\max_z |F_n(z) - \Phi(z)| \longrightarrow 0$, again at a rate of about $c/\sqrt{n}$.

A rule of thumb: the normal approximation is usually adequate when $n \geq 30$.

Often a much smaller n, say $n = 5$, is already quite adequate.

It all depends on what is meant by "adequate."

When $|z|$ is large we have $\Phi(z) \approx 0$ or ≈ 1 and the same will hold for F_n.
Then the relative errors $|F_n(z) - \Phi(z)|/F_n(z)$ or $|F_n(z) - \Phi(z)|/(1 - F_n(z))$
may be more relevant.

Measurement Example

Nuclear magnetic resonance (NMR) spectroscopy is used to measure the distance between nearby hydrogen atoms.

Known: The expected value of this measurement is the actual distance (no bias) the standard deviation is $\sigma = 0.5$ angstroms.

If the measurement process is repeated 36 times, what is the chance that the average measured value $\bar{X}_{36}$ falls within 0.1 angstrom of the true value μ?

$$
\begin{aligned}
P(\mu - 0.1 < \bar{X}_{36} < \mu + 0.1) &= P(\mu - 0.1 - \mu < \bar{X}_{36} - \mu < \mu + 0.1 - \mu) \\
= P(-0.1 < \bar{X}_{36} - \mu < 0.1) &= P\left(\frac{-0.1}{\sigma/\sqrt{n}} < \frac{\bar{X}_{36} - \mu}{\sigma/\sqrt{n}} < \frac{0.1}{\sigma/\sqrt{n}}\right) \\
P\left(-0.1/(0.5/6) < Z_n < 0.1/(0.5/6)\right) &= P\left(-1.2 < Z_n < 1.2\right) \\
\approx P(-1.2 < Z < 1.2) &= \Phi(1.2) - \Phi(-1.2) \\
= \texttt{pnorm}(1.2) - \texttt{pnorm}(-1.2) &= 0.7698607
\end{aligned}
$$

Measurement Example (continued)

$$\text{CLT} \implies \sum_{i=1}^{n} X_i \approx \mathcal{N}(n\mu, n\sigma^2) \quad \text{and} \quad \bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$$

$X \sim \mathcal{D}(\mu, \sigma^2)$ means that X has some distribution with mean μ and variance σ^2.

If someone else independently replicates the previous experiment 64 times,

what is the chance that the two averages are within 0.1 angstroms of each other?

$$X_1, \ldots, X_{36} \sim \mathcal{D}(\mu, \sigma^2) \implies \bar{X}_{36} \approx \mathcal{N}(\mu, \sigma^2/36)$$
$$Y_1, \ldots, Y_{64} \sim \mathcal{D}(\mu, \sigma^2) \implies \bar{Y}_{64} \approx \mathcal{N}(\mu, \sigma^2/64)$$

$$\implies \bar{X}_{36} - \bar{Y}_{64} = \bar{X}_{36} + (-\bar{Y}_{64}) \sim \mathcal{N}(\mu + (-\mu), \sigma^2/36 + \sigma^2/64)$$

$$= \mathcal{N}(0, \sigma^2/36 + \sigma^2/64) = \mathcal{N}(0, 0.25 \cdot (64+36)/(6^2 \cdot 8^2)) = \mathcal{N}(0, 5^2/48^2)$$

$$P(-0.1 < \bar{X}_{36} - \bar{Y}_{64} < 0.1) = P\left(\frac{-0.1}{5/48} < \frac{\bar{X}_{36} - \bar{Y}_{64}}{5/48} < \frac{0.1}{5/48}\right)$$

$$= \Phi(0.96) - \Phi(-0.96) = \texttt{pnorm}(.96) - \texttt{pnorm}(-.96) = 0.6629448$$

More General CLT

In our previous CLT we required the summands X_i to be identically distributed.

Theorem: Let X_i be independent random variables with respective finite means μ_i and variances σ_i^2, $i = 1, \ldots, n$.

Under additional (technical) assumptions of which the following is most relevant

$$\frac{\max(\sigma_1^2, \ldots, \sigma_n^2)}{\sigma_1^2 + \ldots + \sigma_n^2} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty \tag{1}$$

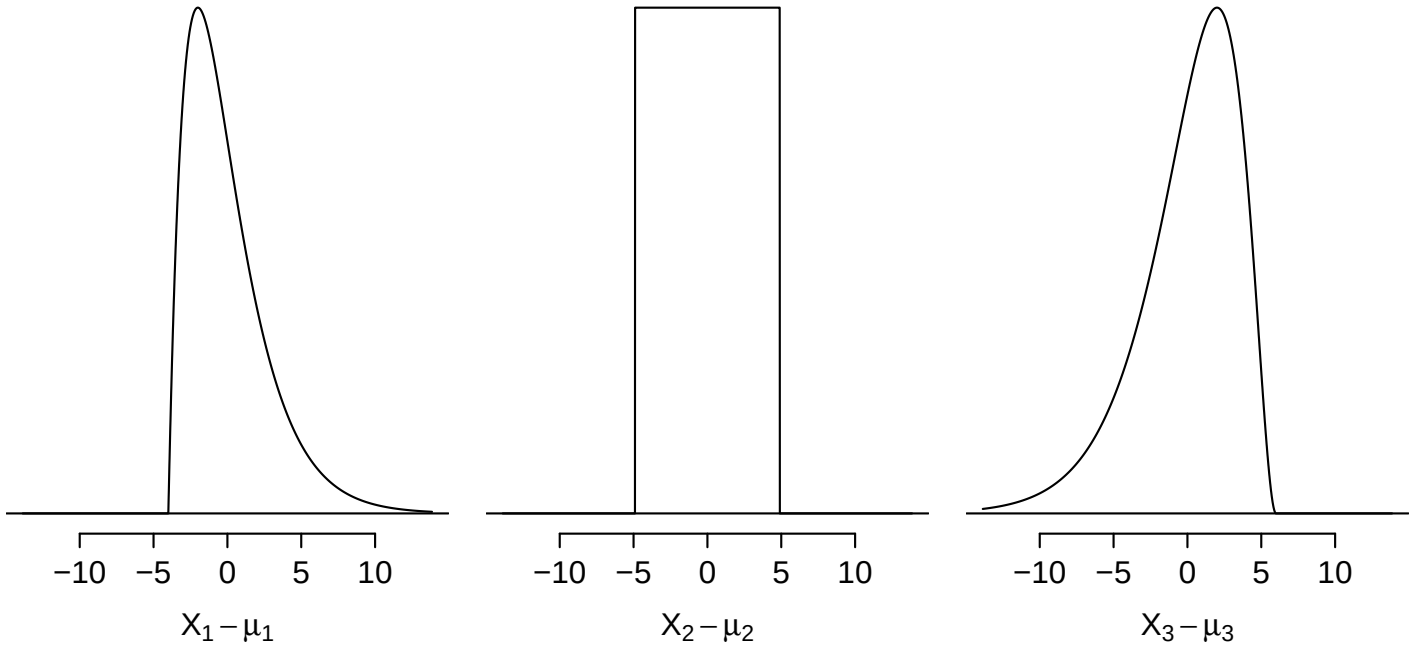
we get that the standardized sum

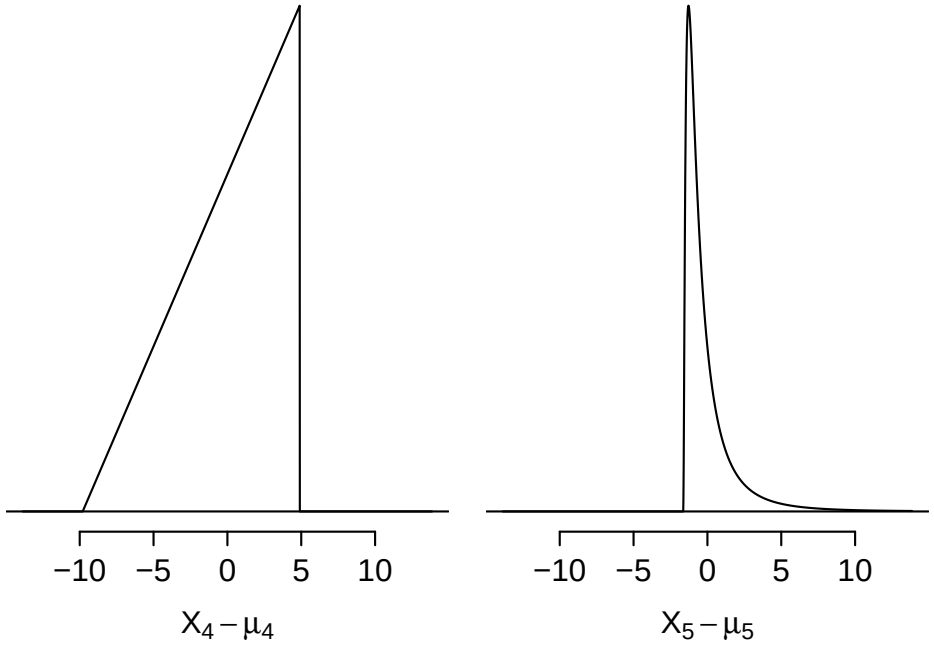
$$Z_n = \frac{X_1 + \ldots + X_n - (\mu_1 + \ldots + \mu_n)}{\sqrt{\sigma_1^2 + \ldots + \sigma_n^2}}$$

has cdf F_n such that

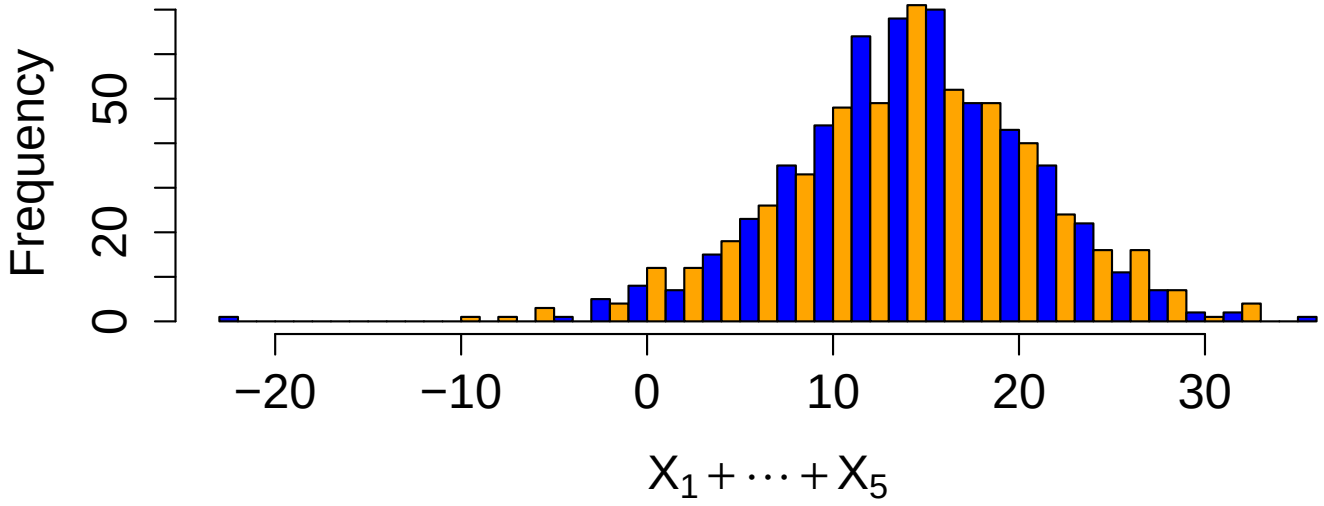
$$P(Z_n \leq z) = F_n(z) \longrightarrow \Phi(z) \qquad \text{as} \quad n \longrightarrow \infty$$

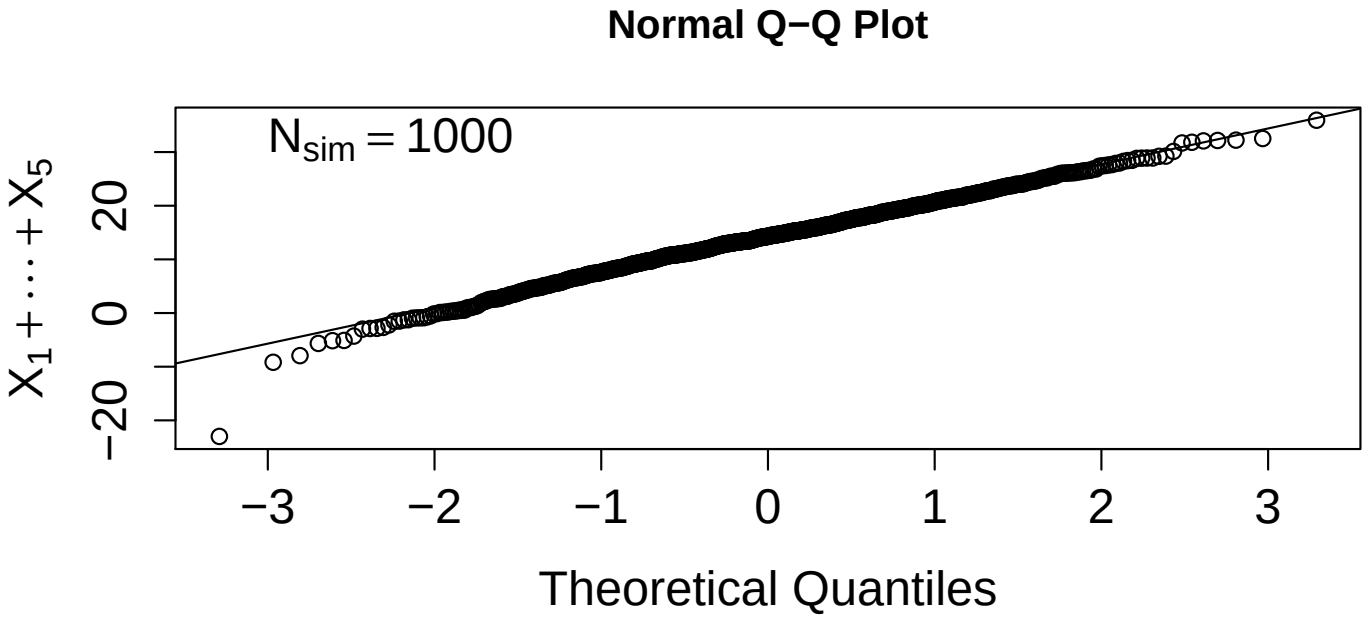
Sampled Densities





CLT in Non-IID Case, $n = 5$





Comments

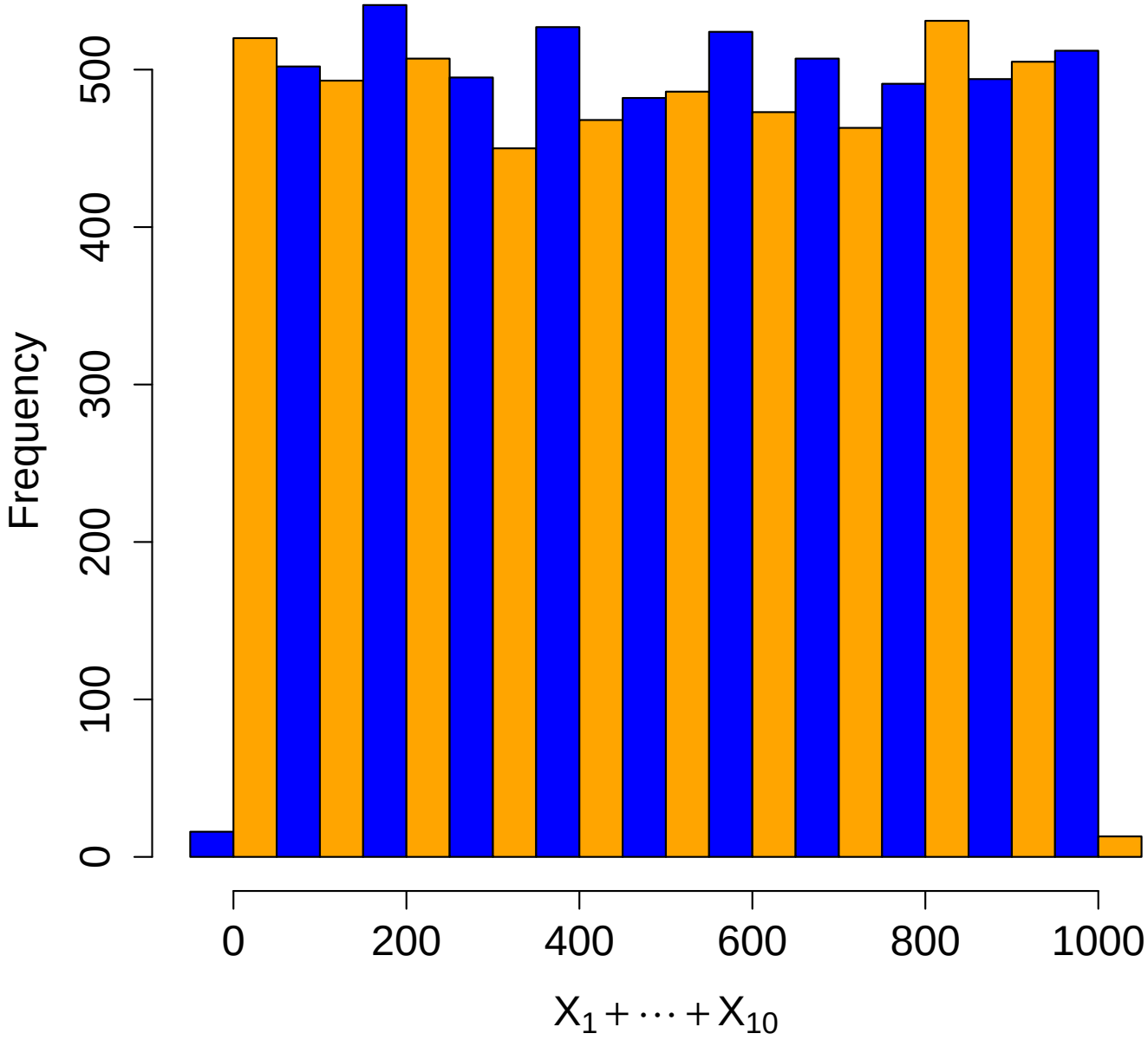
The variance condition (1) makes sure that none of the variances dominate.

All the variances contribute relatively small amounts to the total variability.

For example, if $X_1 \sim \text{Uniform}(0, 1000)$ with a very large variance and all the other random variables $X_i \sim \mathcal{N}(0, 1),\ i = 2, \ldots, n$, then for n not so large the sum $X_1 + \ldots + X_n$ will not be well approximated by a normal distribution,

but will inherit mainly the uniform distribution character of X_1 (see next slide).

$X_1 \sim \text{Uniform}(0, 1000)$ & $X_i \sim \mathcal{N}(0, 1),\ i = 2, \ldots, 10$



Further Comments on the CLT

The initial version of the CLT in the iid case is useful in many situations when an experiment is repeated independently many times and we consider the average $\bar{X}_n$ as or main focus of interest.

The broader non-iid version of the CLT is very useful it rationalizing or modeling a normal distribution for random variables X_i observed in experiments.

This rationalization consists in probing to what extent X_i can be viewed as the sum of many random effects that act more or less independently.

For example, the time to complete a task can be viewed as the sum of the random times to complete many subtasks into which the main task can be decomposed.

Any measurement can be affected by many different sources of small errors.

A Slight Extension of the CLT

Theorem: Let $X_1, X_2 \ldots$ be a sequence of iid random variables with finite mean μ and finite variance σ^2. Suppose that $D_1, D_2, \ldots$ is a sequence of random variables such that $D_n^2 \xrightarrow{P} \sigma^2$ as $n \longrightarrow \infty$ and let

$$T_n = \frac{\bar{X}_n - \mu}{D_n/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{D_n}$$

Then for any $t \in R$ we have

$$F_n(t) = P(T_n \leq t) \longrightarrow \Phi(t) \qquad \text{as} \qquad n \longrightarrow \infty$$

Note that D_n/σ and its reciprocal basically behave like the constant 1 as $n \longrightarrow \infty$.

In our previous measurement example we assumed a known $\sigma = 0.5$ angstrom.

Typically σ is not known, but one can get an estimate of σ^2,

say the plug-in sample estimate $\widehat{\sigma}_n^2$.

$$\widehat{\sigma}_n^2 \xrightarrow{P} \sigma^2$$

Recall

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$$

By the WLLN applied to the averages of the X_i and X_i^2

$$\bar{X}_n \xrightarrow{P} \mu \implies \bar{X}_n^2 \xrightarrow{P} \mu^2 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} E(X_i^2)$$

$$\implies \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 \xrightarrow{P} E(X_i^2) - \mu^2 = \sigma^2$$

While the above sequence of conclusions still require some technical details, our understanding of $\xrightarrow{P}$ should make them quite evident.