The Central Limit Theorem
The Bridge Between Probability and Statistics

Andrew J. Spieker, PhD

Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

Outline

The Central Limit Theorem

- Normal approximation to the binomial distribution
- Sampling distributions

Outline

Beware!

- If you've taken a statistics course, you *may* have heard of a law that says something about things becoming approximately normally distributed as sample sizes get large.
- **Forget what you think you've heard about that rule!**
  - We are about to learn it the *right* way!

Review of Bernoulli distribution

Recall the Bernoulli distribution (binary variables)

- The prevalence of some characteristic or trait (e.g., BRCA mutation) in the population is p.
- Randomly sample a single individual from that population; let X be 1 if that person has this trait, and 0 otherwise.
- Then, $X \sim \text{Bernoulli}(p)$, so that X takes on the value 1 with probability p and the value 0 with probability $1 - p$.

Review of binomial distribution

Recall

- The prevalence of some characteristic or trait (e.g., BRCA mutation) in the population is p.
- Now, sample n people and record 0/1 for each depending on whether they have the trait. Each of those n random variables, $X_1, \ldots, X_n$, has a Bernoulli(p) distribution.
- So, $T = \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n \sim \text{Binomial}(n, p)$.
- **Two key points**:
  1. T is the **sum** of n *independent and identically distributed* (iid) random variables.
  2. Obtaining a value for T can be thought of as conducting a single study, *or* conducting n independent "mini-studies."

Review of binomial distribution

Limiting behavior

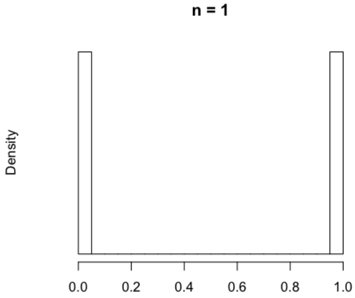
- So, suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ represent n *independent and identically distributed* random variables (e.g., the result of sampling people from a population with mutation prevalence p).
- $T = \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n \sim \text{Binomial}(n, p)$.
  - Randomly sample n individuals from the target population and *count* the number of those n who have the trait/characteristic; that count has a binomial distribution.
- It turns out that if n is very large, then T has an approximate normal distribution.
  - Let's take a look!

Example: Prostate cancer cells

Binomial distribution!

- The Binomial(n, p) distribution has two parameters (n and p)!
- Let us imagine that 50% of men over the age of 60 would test positive for the presence of prostate cancer cells ($p = 0.5$).
- We want to see what happens when we sample n people from this population and count the number of men who test positive for the presence of prostate cancer cells.

Binomial distribution: Sample one individual

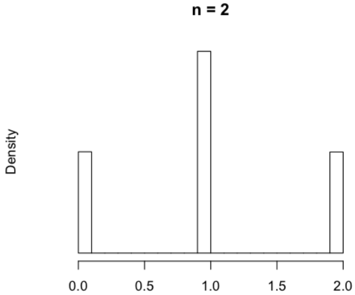


Binomial($n = 1, p = 0.5$)

(Half the time, you would sample an individual with prostate cancer cells, and half the time, you would sample an individual without.)

Binomial distribution: Sample two individuals

How many have prostate cancer cells?

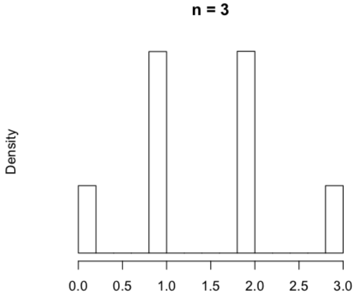


Binomial($n = 2, p = 0.5$)

(25% of the time, neither would; 50% of the time, exactly one would; and 25% of the time, both would.)

Binomial distribution: Sample three individuals

How many have prostate cancer cells?

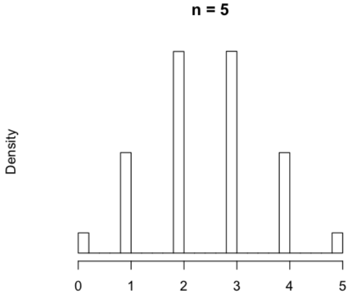


Binomial$(n = 3, p = 0.5)$

(12.5% of the time, none; 37.5% of the time, exactly one; 37.5% of the time, exactly two; and 12.5% of the time, all three.)

Binomial distribution: Sample five individuals

How many have prostate cancer cells?

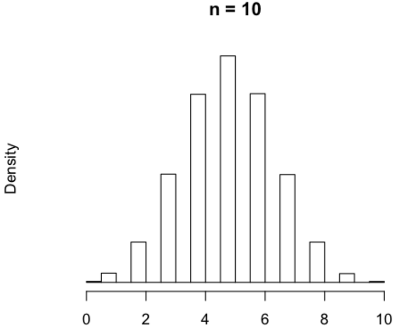


Binomial($n = 5, p = 0.5$)

(. . . and so on. . .)

Binomial distribution: Sample ten individuals

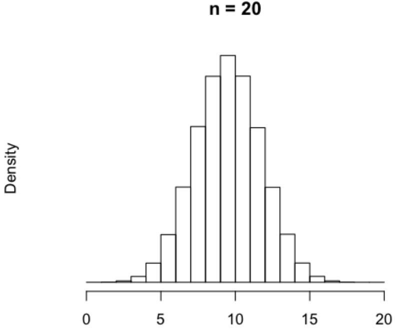
How many have prostate cancer cells?



Binomial($n = 10$, $p = 0.5$)

Binomial distribution: Sample twenty individuals

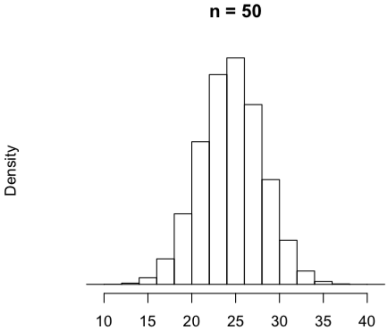
How many have prostate cancer cells?



Binomial$(n = 20, p = 0.5)$

Binomial distribution: Sample fifty individuals

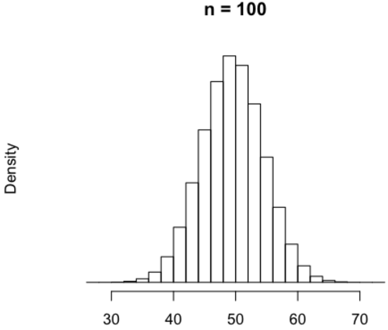
How many have prostate cancer cells?



Binomial$(n = 50, p = 0.5)$

Binomial distribution: Sample one-hundred individuals

How many have prostate cancer cells?



Binomial(n = 100, p = 0.5)

Binomial distribution

Thoughts?

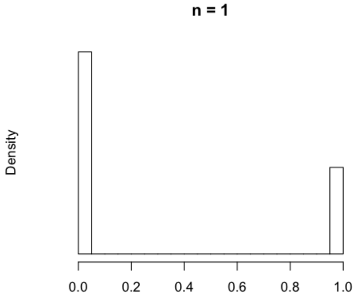
- In this example, $\text{Binomial}(n, p = 0.5)$ is always symmetric, regardless of what n is.
- I think we can all agree that as n gets larger, appears more like a normal distribution.
- But *which* normal distribution? Suppose $n = 100$:
  - In truth, $E[T] = np = 50$ and $\text{Var}[T] = np(1 - p) = 25$: one *hopes* that if T "looks normal," that it would look like the normal distribution of mean 50 and variance 25.
  - It does!
- Let's make sure Andrew isn't just making things up and try another example.

Example: Hypertension

Binomial distribution!

- The Binomial(n, p) distribution has two parameters (n and p)!
- Let us imagine that 30% of people over the age of 50 suffer from hypertension ($p = 0.3$).
- We want to see what happens when we sample n people from this population and count the number of people who have hypertension.

Binomial distribution: Sample one individual

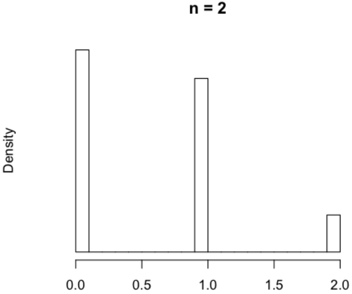


Binomial($n = 1, p = 0.3$)

(30% of the time, would sample individual with hypertension; 70% of the time, would sample individual without.)

Binomial distribution: Sample two individuals

How many have hypertension?

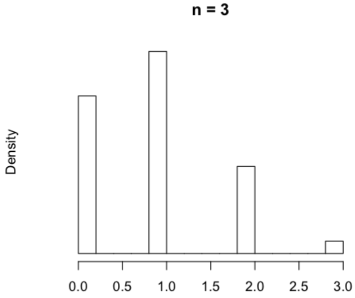


Binomial($n = 2, p = 0.3$)

(49% of the time, neither would; 42% of the time, exactly one would; and 9% of the time, both would.)

Binomial distribution: Sample three individuals

How many have hypertension?

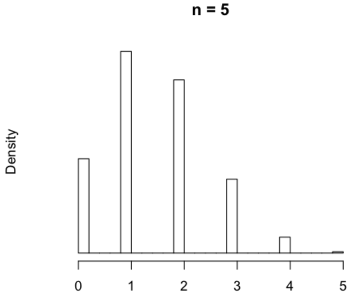


Binomial($n = 3, p = 0.3$)

(34.3% of the time, none; 44.1% of the time, exactly one; 18.9% of the time, exactly two; and 2.7% of the time, all three.)

Binomial distribution: Sample five individuals

How many have hypertension?

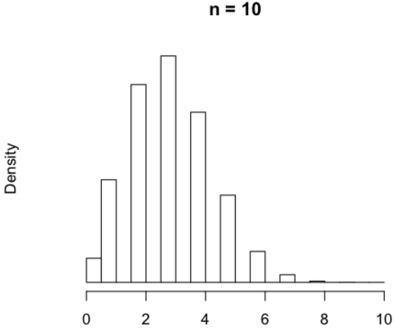


Binomial$(n = 5, p = 0.3)$

(...and so on...)

Binomial distribution: Sample ten individuals

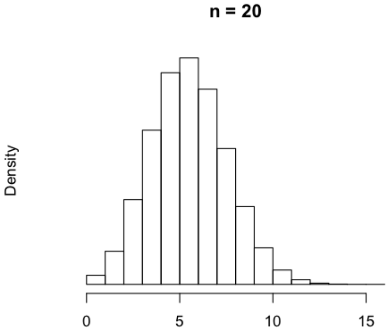
How many have hypertension?



Binomial($n = 10, p = 0.3$)

Binomial distribution: Sample twenty individuals

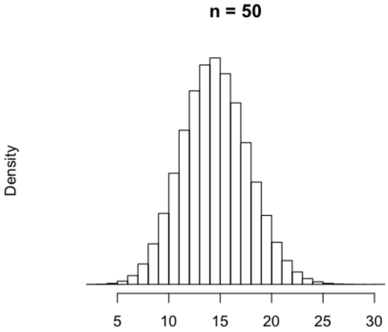
How many have hypertension?



Binomial($n = 20, p = 0.3$)

Binomial distribution: Sample fifty individuals

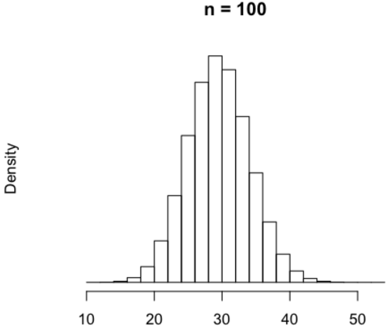
How many have hypertension?



Binomial($n = 50$, $p = 0.3$)

Binomial distribution: Sample one-hundred individuals

How many have hypertension?



Binomial($n = 100, p = 0.3$)

Binomial distribution

Thoughts?

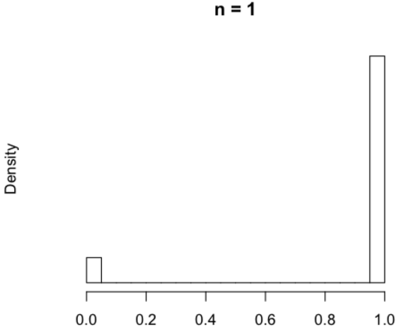
- In this example, Binomial$(n, p = 0.3)$ is not generally symmetric; generally, appears to be a bit right-skewed with lower sample sizes.
- As n gets larger, the distribution becomes more symmetric. In fact, it starts to look like a normal distribution!
- But *which* normal distribution? Suppose $n = 100$:
  - In truth, $\mathrm{E}[T] = np = 30$ and $\mathrm{Var}[T] = np(1-p) = 21$: one *hopes* that if T "looks normal," that it would look like the normal distribution of mean 30 and variance 21.
  - It does!
- Just one more time? Let's go wild and try $p = 0.9$.

Example: Smoking and lung cancer

Binomial distribution!

- The Binomial(n, p) distribution has two parameters (n and p)!
- Let us imagine that 90% of people under the age of 60 with small-cell lung cancer are smokers ($p = 0.9$).
- We want to see what happens when we sample n people from this population and count the number of smokers.

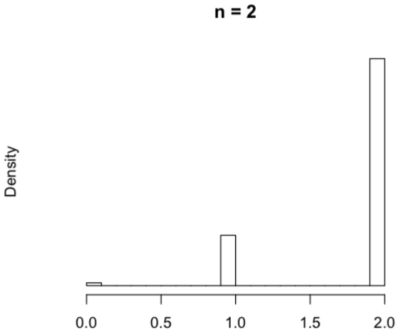
Binomial distribution: Sample one individual



Binomial($n = 1, p = 0.9$)

(90% of the time, would sample a smoker; 10% of the time, would sample non-smoker.)

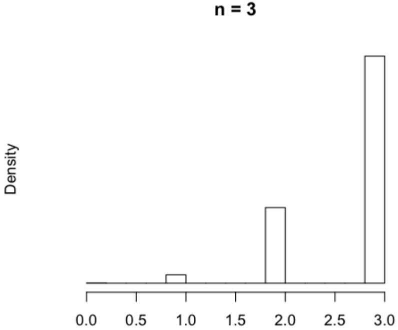
Binomial distribution: Sample two individuals



Binomial($n = 2, p = 0.9$)

(1% of the time, neither would; 18% of the time, exactly one would; and 81% of the time, both would.)

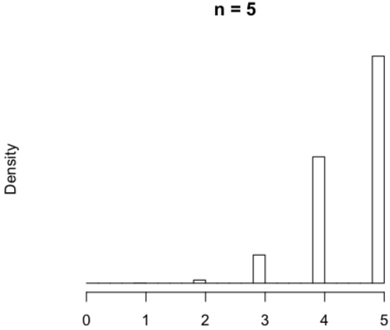
Binomial distribution: Sample three individuals



Binomial($n = 3, p = 0.9$)

(0.1% of the time, none; 2.7% of the time, exactly one; 24.3% of the time, exactly two; and 72.9% of the time, all three.)

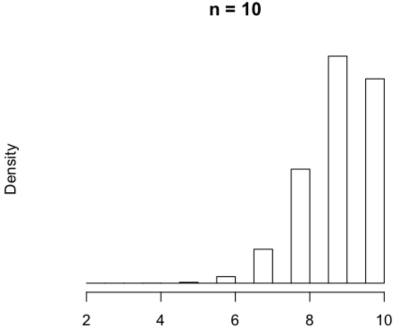
Binomial distribution: Sample five individuals



Binomial($n = 5, p = 0.9$)

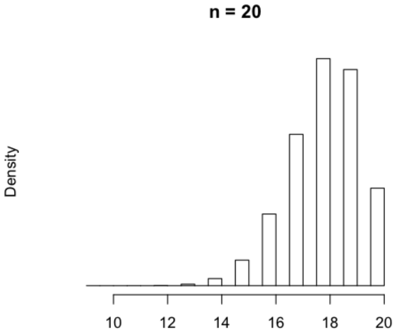
(. . . and so on. . .)

Binomial distribution: Sample ten individuals



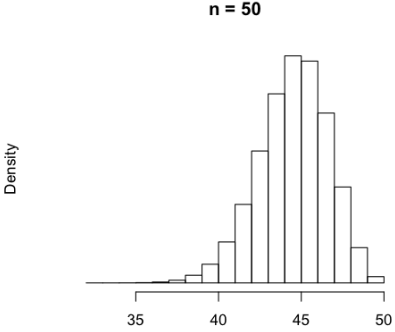
$$\text{Binomial}(n = 10, p = 0.9)$$

Binomial distribution: Sample twenty individuals



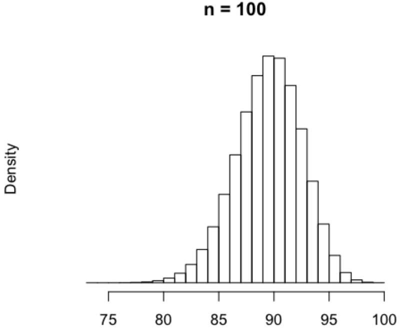
Binomial(n = 20, p = 0.9)

Binomial distribution: Sample fifty individuals



Binomial($n = 50, p = 0.9$)

Binomial distribution: Sample one-hundred individuals



Binomial($n = 100$, $p = 0.9$)

Binomial distribution

Thoughts?

- In this example, Binomial($n, p = 0.9$) is not generally symmetric; generally, appears to quite left-skewed with lower sample sizes.
- As n gets larger, the distribution becomes more symmetric. In fact, it starts to look like a normal distribution!
- But *which* normal distribution? Suppose $n = 100$:
  - In truth, $\mathrm{E}[T] = np = 90$ and $\mathrm{Var}[T] = np(1-p) = 9$: one *hopes* that if T "looks normal," that it would look like the normal distribution of mean 90 and variance 9.
  - It does!

The statement

What's going on?

- It turns out that for large sample sizes, the Binomial distribution can be approximated by a normal distribution.
- Specifically, if $X \sim \text{Binomial}(n, p)$, if n is large enough, then:

$$X \dot{\sim} \mathcal{N}(np, np(1-p))$$

- Recall: We use symbol "$\dot{\sim}$" to mean "is approximately distributed as," as opposed to the symbol "$\sim$", which means "is exactly distributed as."

Normal approximation to the binomial distribution

Example: Stage II bladder cancer

- For Stage II bladder cancer, the 5-year relative survival rate is approximately 63%
- You randomly sample 250 individuals with Stage-II bladder cancer and, at the end of five years, determine the number who are still alive. Let X denote the number still alive.
- Therefore, $X \sim \text{Binomial}(n = 250, p = 0.63)$.
- Why would we not want to compute $\text{P}(X > 150)$ "by hand"?
  - Because, we don't want to have to evaluate the probability mass function 151 times.
- Exercise: Use the normal approximation to the binomial distribution to approximate $\text{P}(X > 150)$.

Normal approximation to the binomial distribution

Example: Stage II bladder cancer

- Here, $X \sim \text{Binomial}(n = 250, p = 0.63)$.
- We want to compute $\text{P}(X > 150)$.
  - $X \dot{\sim} \mathcal{N}(\mu = 250 \times 0.63, \sigma^2 = 250 \times 0.63 \times 0.37)$.
  - $X \dot{\sim} \mathcal{N}(\mu = 157.5, \sigma^2 = 58.275)$.
  - Recall: $Z = (X - 157.5)/\sqrt{58.275} \dot{\sim} \mathcal{N}(0, 1)$.
  - $\text{P}(Z > (150 - 157.5)/\sqrt{58.275}) = \text{P}(Z > -0.928) = 0.837$.
  - NB: The true answer, using statistical software, is 0.821, not very far off from our approximation. The higher the sample size, the better the approximation.

The reason

Key point

- Binomial dist. can be approximated by normal dist. of the same mean and variance if sample size is large.
- Phenomenon doesn't occur for just *any* random variable!
- So, why does this happen in *this* case?
  - Because a binomial random variable can be expressed as a *sum* of independent, identically distributed (iid) random variables.
- Specifically, $T = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Bernoulli}(p)$.
- There is a statistical "rule" that says that sums of *iid* random variables will tend to have an approximate normal distribution the sample size is large (central limit theorem).

Normal approximations to sums

The central limit theorem

- Suppose $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2.
- Then, if n is "large enough," then:

$$T = \sum_{i=1}^{n} X_i \mathrel{\dot\sim} \mathcal{N}(n\mu, n\sigma^2)$$

- This is true *even* when the individual X's are not themselves sampled from a normal distribution! That's the magic of the theorem.

Other facts

Normal approximation to negative binomial distribution

- If $X \sim \text{NegBinomial}(k, p)$, and if k is large enough:

$$X \mathrel{\dot\sim} \mathcal{N}\left(\mu = \frac{k}{p}, \sigma^2 = \frac{k(1-p)}{p^2}\right)$$

- Why does this happen? Because X can be expressed as the sum of k *iid* Geometric(p) random variables!

Other facts

Normal approximation to Poisson distribution

- If $X \sim \text{Poisson}(\lambda)$, and if λ is large enough:

$$X \dot{\sim} \mathcal{N}\left(\mu = \lambda, \sigma^2 = \lambda\right)$$

- Why does this happen? Because X can be expressed as the sum of n *iid* Poisson(λ/n) random variables!

Normal approximations to distributions other than the binomial distritbuion often appear as optional problems on homework. ☺

How else is this useful?

Using the central limit theorem

- Sample $n = 100$ people and record their LDL values.
- These LDL values are random variables: each takes on a single value as result of a random sampling process.
- If X_i denotes the LDL value for subject i, then the *sample mean*, $\overline{X}$ is, too, a random variable:

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{100} X_i$$

- What does it mean for $\overline{X}$ to be a random variable?
- In a single study, $\overline{X}$ takes on a value, $\overline{x}$ (of many possible values), as the result of a random sampling process.
  - Should we do this study again, would get a different value for $\overline{X}$. And again, would get something different from other two.

How else is this useful?

Using the central limit theorem

- If X_i denotes the LDL value for subject i, then the *sample mean*, $\overline{X}$ is, too, a random variable:

$$\overline{X} = \frac{1}{100} \sum_{i=1}^{100} X_i$$

- Point: If central limit theorem tells you that sums of *iid* random variables are approximately normally distributed, then the sample mean is approximately normally distributed.
- Why? Because if, $T = \sum_{i=1}^{n} X_i \dot{\sim} \mathcal{N}(n\mu, n\sigma^2)$, then:

$$\overline{X} = T/n \dot{\sim} \mathcal{N}(\mu, \sigma^2/n).$$

- **I am equally interested in your ability to interpret what this means as I am in your ability to apply this!**

Sampling distribution of the mean

Recall:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean, is a random variable.
  - The sample mean LDL from a study of $n = 100$ people, for example. It takes on one of many, many possible values in a given study.
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, a *statistic*, computed from one specific data set.
  - $\overline{x}$ denotes the specific value of the sample mean computed from your single study of, for example, $n = 100$ people. $\overline{x} = 120$ μg/dL, for instance.

Sampling distribution of the mean

For clarity:

- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a random variable with a distribution.
- $\overline{x}$ is computed from a single study, and takes on one value of *many* possible values it could have taken on.

Study number (k)	$\overline{X}$
$k = 1$	$\overline{x}_1$
$k = 2$	$\overline{x}_2$
$k = 3$	$\overline{x}_3$
$\vdots$	$\vdots$

Sampling distribution of the mean

Applying the central limit theorem

- Let $X_1, \ldots, X_n$ the LDL values for n randomly sampled individuals (assume n is large).
- If I were to conduct the above study in the same way, repeatedly, each time recording the sample mean, what would the distribution of those sample means look like?
- **Answer**:

$$\overline{X} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

- The beauty here is that we don't even need to know the distribution of LDL! For large enough samples, we know (approximately) the distribution of the sample means.

Mean and variance of sample mean

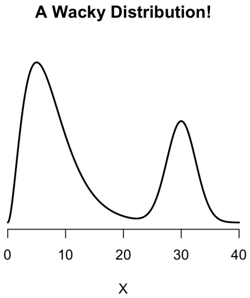
Recall:

- Some math in an earlier set of lecture notes showed us that:
  1. The sample mean, $\overline{X}$, is *unbiased* for the population mean, μ.
  2. The sample mean, $\overline{X}$ gets closer and closer to μ the larger your sample size, n, becomes larger.
- You're not expected to remember or replicate the math (it's there for your reference), but the two concepts above are important to understand.
- **Discussion point**: How does the central limit theorem square with the two points above?

Sampling distribution of the mean

Example: Something more wacky?

- Sampling a wacky continuous distribution–say, net insurance claims (in thousands of dollars).

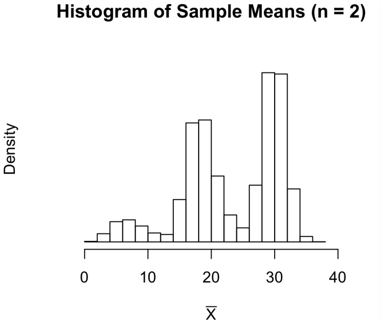


- NB: If $n = 1$, then the sampling distribution of $\overline{X}$ is the distribution of X (make sure this makes sense)!

Sampling distribution of the mean

Example: Something more wacky?

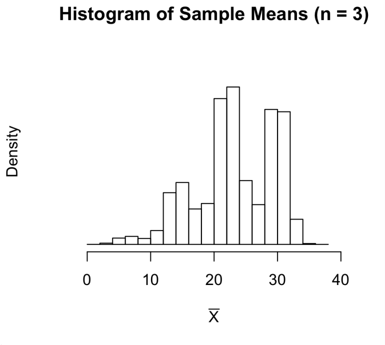
- Example: Suppose $X_1, X_2 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

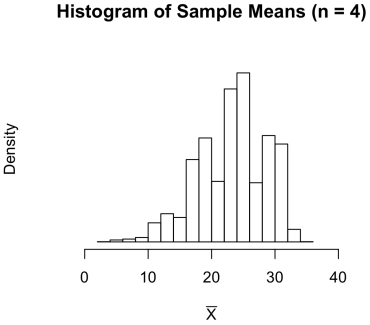
- Example: Suppose $X_1, X_2, X_3 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

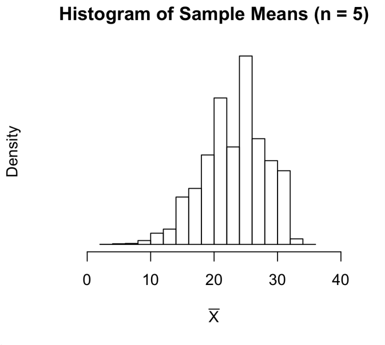
- Example: Suppose $X_1, \ldots, X_4 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

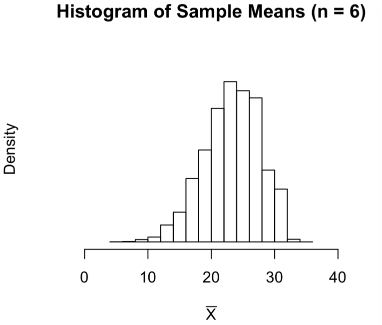
- Example: Suppose $X_1, \ldots, X_5 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

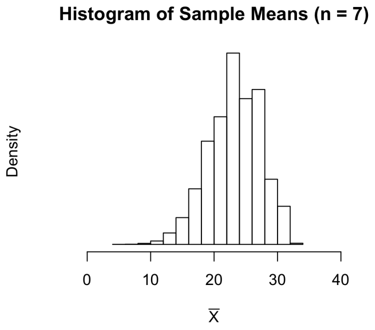
- Example: Suppose $X_1, \ldots, X_6 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

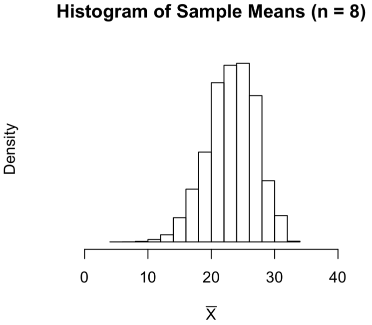
- Example: Suppose $X_1, \ldots, X_7 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

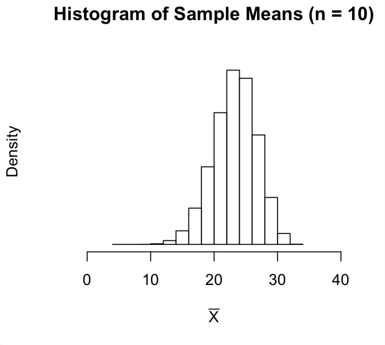
- Example: Suppose $X_1, \ldots, X_8 \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

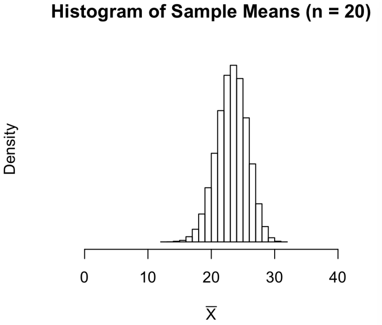
- Example: Suppose $X_1, \ldots, X_{10} \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

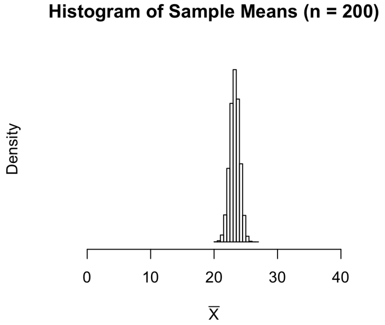
- Example: Suppose $X_1, \ldots, X_{20} \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

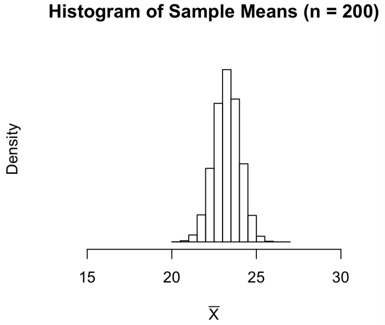
- Example: Suppose $X_1, \ldots, X_{200} \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



Sampling distribution of the mean

Example: Something more wacky?

- Example: Suppose $X_1, \ldots, X_{200} \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?

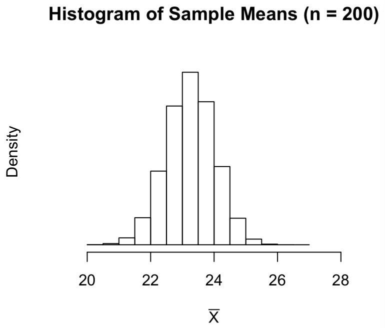


- Same figure, but reducing the x-axis range.

Sampling distribution of the mean

Example: Something more wacky?

- Example: Suppose $X_1, \ldots, X_{200} \sim$ Wacky Distribution. What is the *sampling* distribution of $\overline{X}$?



- Same figure, but reducing the x-axis range further.

Sampling distribution of the mean

Believe me yet? ☺

- It really does seem that, even if the distribution of your variable of interest is skewed and/or multimodal (or otherwise wacky), whether it is:
    - LDL
    - Insurance claims
    - Blood pressure
    - Cognitive abilities screening instrument
    - Number of hours in ICU

  ... there is **nothing** you can do to escape the following fact:
- **If your sample size is large enough, the distribution of the sample means over study replicates will be approximately normally distributed.**
    - In particular, with mean μ and variance σ^2/n.

Sampling distribution of the mean

Understanding the theorem:

- Sample $X_1, \ldots, X_n$ for $n = 1$. In this case, $\bar{x} = x_1$.
- If I repeated this study "infinitely" many times, what would the distribution of $\overline{X}$ look like?

Study number (k)	$n = 1$
$k = 1$	$\bar{x}_1$
$k = 2$	$\bar{x}_2$
$k = 3$	$\bar{x}_3$
$\vdots$	$\vdots$
Dist. of $\overline{X}$	Same as dist. of X

Sampling distribution of the mean

Understanding the theorem:

- Sample $X_1, \ldots, X_n$ for $n =$ "small".
- If I repeated this study "infinitely" many times, what would the distribution of $\overline{X}$ look like?

Study number (k)	$n =$ "small"
$k = 1$	$\overline{x}_1$
$k = 2$	$\overline{x}_2$
$k = 3$	$\overline{x}_3$
$\vdots$	$\vdots$
Dist. of $\overline{X}$	???

Sampling distribution of the mean

Understanding the theorem:

- Sample $X_1, \ldots, X_n$ for $n =$ "medium".
- If I repeated this study "infinitely" many times, what would the distribution of $\overline{X}$ look like?

Study number (k)	$n =$ "medium"
$k = 1$	$\overline{x}_1$
$k = 2$	$\overline{x}_2$
$k = 3$	$\overline{x}_3$
$\vdots$	$\vdots$
Dist. of $\overline{X}$	Closer to normal than small sample

Sampling distribution of the mean

Understanding the theorem:

- Sample $X_1, \ldots, X_n$ for $n =$ "huge".
- If I repeated this study "infinitely" many times, what would the distribution of $\overline{X}$ look like?

Study number (k)	$n =$ "huge"
$k = 1$	$\overline{x}_1$
$k = 2$	$\overline{x}_2$
$k = 3$	$\overline{x}_3$
$\vdots$	$\vdots$
Dist. of $\overline{X}$	Approximately normal*

*Theorem asserts that there is an n large enough that this will be the case.

The central limit theorem

Formally: The central limit theorem!

- Suppose $X_1, \ldots, X_n$ are independently sampled from a common distribution with mean μ and variance σ^2. As n grows larger and larger,

$$\mathrm{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < s\right) \longrightarrow \mathrm{P}(Z < s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s} e^{-\frac{1}{2}t^2} dt,$$

  ... where $Z \sim \mathcal{N}(0, 1)$.

- **You do not need to remember/use this formula.**

The central limit theorem

Ways of stating the central limit theorem!

- Suppose $X_1, \ldots, X_n$ are *iid* with common mean μ and variance σ^2. Then, for large sample sizes, n:

$$
\begin{aligned}
\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad &\dot{\sim} \quad \mathcal{N}(0, 1), \text{ or} \\
\sqrt{n}(\overline{X} - \mu) \quad &\dot{\sim} \quad \mathcal{N}(0, \sigma^2), \text{ or} \\
\overline{X} \quad &\dot{\sim} \quad \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)
\end{aligned}
$$

The central limit theorem

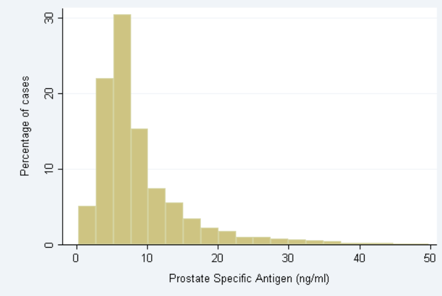
Thoughts

- To me, this is wild for two reasons:
  1. This works for any distribution, whether discrete or continuous. Whether symmetric or skewed. Whether the range is finite or infinite. Whether it's unimodal or bimodal.
  2. How lucky we are that the sample mean looks more and more like a *normal* distribution–one that we understand so well.
- Caution: The central limit theorem does not say anything about the distribution of the variable itself as the sample size grows. The distribution of the variable you're measure *does not change* with sample size.

The central limit theorem

Application: PSA

- PSA (prostate specific antigen) a biomarker used to detect prostate cancer.
- Among men undergoing surgery for prostate cancer:
  - Mean: approximately 10 ng/mL.
  - Variance: approximately 11 ng/mL.



The central limit theorem

Application: PSA

- Among men undergoing surgery for prostate cancer:
  - Mean: approximately 10 ng/mL.
  - Variance: approximately 11 ng/mL.
- **Exercises**: Sample $n = 120$ men undergoing surgery for prostate cancer, and record their PSA values.
  - You decide to plot a histogram of their PSA values. Do you believe that histogram would suggest that the distribution of the PSA values would be approximately normally distributed?
  - You compute the sample mean PSA value. With approx. what probability would it be greater than 11 ng/mL?
  - Approximately what values mark the 2.5th and 97.5th percentiles of the sampling distribution of $\overline{X}$?

The central limit theorem

Application: PSA

- True mean PSA: approximately 10 ng/mL.
- True variance of PSA: approximately 11 ng/mL.
- Sample $n = 120$ men undergoing surgery for prostate cancer, and record their PSA values.
- **Exercise**: You decide to plot a histogram of their PSA values. Do you believe that histogram would suggest that the distribution of the PSA values would be approximately normally distributed?
    - **Answer**: No!

The central limit theorem

Application: PSA

- True mean PSA: approximately 10 ng/mL.
- True variance of PSA: approximately 11 ng/mL.
- Sample $n = 120$ men undergoing surgery for prostate cancer, and record their PSA values.
- **Exercise**: You compute the sample mean PSA value. With approx. what probability would it be greater than 10.3 ng/mL?
  - **Answer**: The central limit theorem asserts that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\cdot}{\sim} \mathcal{N}(0,1).$$

  - Here, $z = (10.3 - 10)/(\sqrt{11/120}) = 0.9909$.
  - $\mathrm{P}(Z > 0.9909) = 1 - 0.8391 = 0.161$.

The central limit theorem

Application: PSA

- True mean PSA: approximately 10 ng/mL.
- True variance of PSA: approximately 11 ng/mL.
- Sample $n = 120$ men undergoing surgery for prostate cancer, and record their PSA values.
- **Exercise**: Approximately what values mark the 2.5th and 97.5th percentiles of the sampling distribution of $\overline{X}$?
  - **Answer**: The central limit theorem asserts that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \mathrel{\dot\sim} \mathcal{N}(0,1).$$

  - Recall: ± 1.96 mark the 2.5th and 97.5th percentiles of standard normal dist. Can convert back to the PSA scale:
  - $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}} = 10 \pm 1.96 \times \sqrt{11/120} = [9.41 \text{ ng/mL}, 10.6 \text{ ng/mL}]$.

The central limit theorem

Application: PSA

- Interpretation of activity results: If I were to conduct this study over, and over, and over, each time recording the sample mean, then:
  - 16.1% of those sample means would be greater than 10.3 ng/mL.
  - The 2.5th and 97.5th percentiles of the distribution of the sample means are approximately 9.41 ng/mL and 10.6 ng/mL.

The central limit theorem

Aside: Why the "normal" distribution?

- The normal distribution is maximally "disorderly".
  - The *entropy* of a random variable X is $\mathrm{E}[\log X]$.
  - If X is normally distributed with expectation μ and variance σ^2, then *no other* random variable with expectation μ and variance σ^2 has higher entropy.
- Makes sense that if we randomly sample values from a distribution, the repeat-sample distribution of the sample mean should head toward the most *disorderly* distribution.

The central limit theorem: The point?

Keeping our eye on the prize!

- May wish to *estimate*, population mean μ.
  - Know how to do this: compute $\overline{x}$ in the sample.
- Want to quantify degree of precision with which we know μ.
  - Rely on information about *sampling* distribution of $\overline{X}$: "what does the distribution of $\overline{X}$ look like if I repeat this study?"
    1. Typically, don't know *exact* sampling dist. of $\overline{X}$, and ...
    2. Typically, cannot actually *perform* study over and over...
- **Central limit theorem (CLT) is our friend!** It is ultimately a statement about the approximate sampling distribution of $\overline{X}$ for large sample sizes. You will use it for the rest of the year, and pretty much the rest of your research career.

Sampling distribution of the mean

Population vs. sample mean

- I know how to compute $\overline{x}$ from a data set.
- With this information, I want to uncover some sort of information about μ, the population mean.
- In our blood pressure example, this could mean asking, for example, the following questions:
  - What are the true values of the population average blood pressure with which my data are consistent?
    - Confidence intervals! Characterize precision of estimate.
  - If the true value of the population mean blood pressure were 130 mm Hg, with what frequency would I observe a sample mean at least as high as the one I observed?
    - p-values! Strength of evidence in support of your hypothesis.

Sampling distribution of the mean

Population vs. sample mean

- Ability to obtain answers to these sorts of questions regarding the population mean is inherently tied to our understanding of the *sampling distribution* of $\overline{X}$.
- That is, we need to ask ourselves the following question:
  - "If I were to complete this study over and over again, and each time compute $\overline{x}$, what would the distribution of the values of $\overline{x}$ look like?"
- Reminder: We typically cannot name the distribution of $\overline{X}$.
  - If we *knew* the exact form of the distribution of $\overline{X}$, then in some sense this wouldn't be too hard of a problem.

The central limit theorem

Summary

- Major theorem!
- Gives us approximate sampling distribution of $\overline{X}$ when the population parameters of X are known.
    - If I repeated the study over and over, recording the value $\overline{x}$ each time, what would the distribution of those values be? With large samples, approximately normal!
- Building foundation to handle the case where the population parameters are *unknown*.
- In turn, we will shortly be able to answer questions about:
    - The precision of our estimate of the population mean.
    - The strength of our evidence for a hypothesis regarding the population mean.

I leave you with spooky statistics!