Class 14

DATA1220-55, Fall 2024

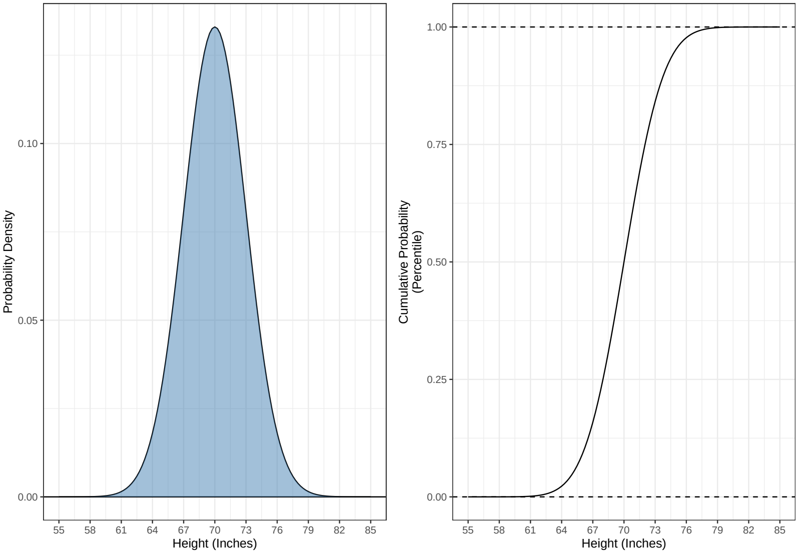
Sarah E. Grabinski

2024-09-30

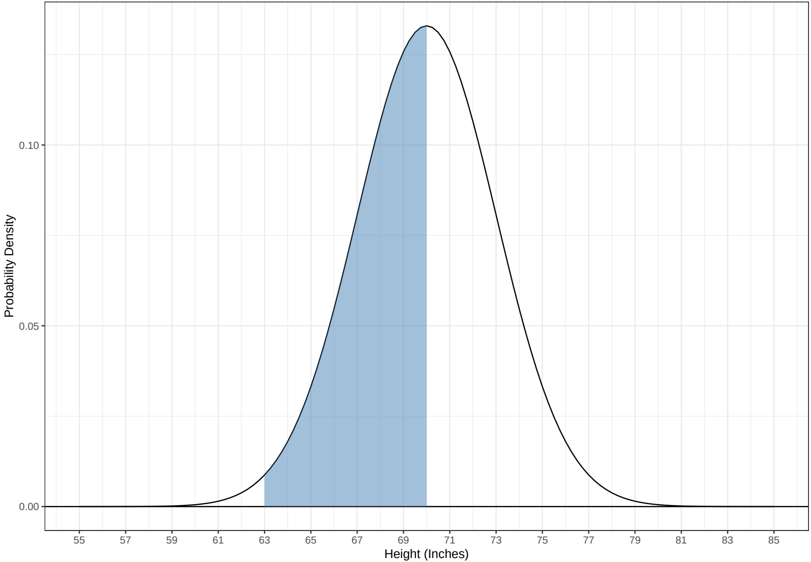
Example: Calculating Probabilities with Normal Distributions

- The average male height is ~70" (5'10") and approximately follows the distribution $N(70, 3)$.
- The average female height is ~63" (5'3") and approximately follows the distribution $N(63, 3)$.
- What is the probability that a random male is taller than the average female but still shorter than the average male?

What's the sample space?



Visualizing the problem



Breaking it down...

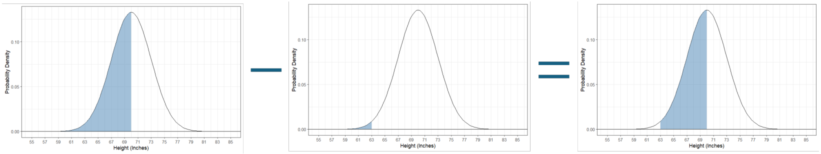


Figure 1: You can think of this problem as the difference between two percentiles.

Calculating the difference in R

The probability that a man is below average in height for men but above average in height for women is the difference between the percentile for 70" and the percentile for 63" in the distribution of male heights or 49%.

```
pnorm(70, mean = 70, sd = 3) - pnorm(63, mean = 70, sd = 3)
```

```
[1] 0.4901847
```

Chapter 5: Foundations for Inference

- 5.1: Point estimates and sampling variability
- 5.2: Confidence intervals for a proportion
- 5.3: Hypothesis testing for a proportion

Random Sampling

Why do we sample?

- Easier, cheaper, convenience, etc.
- Hard to do a census

What are the downsides of sampling?

- May not represent study/target population
- Methods might introduce bias
- Results could be due to chance

How certain are we that our data is representative?

Example: Estimating Proportions

- Bureau of Justice Statistics September 2021 Special Report on Recidivism of Prisoners
- Sampled 73,600 prisoner records from National Corrections Reporting Program on 409,300 state prisoners released across 24 states in 2008, representing 69% of all persons released from state prisons
- 89% of sampled prisoners were male, and 11% of sampled prisoners were female.
- 42.9% of sampled prisoners were rearrested within 1 year. 43.9% of male sampled prisoners were rearrested within 1 year. 34.4% of female sampled prisoners were rearrested within 1 year.

Example: As a dataframe in R

	id	sex	reoffend
62996	prisoner62996	Male	Not Rearrested
60533	prisoner60533	Male	Not Rearrested
33334	prisoner33334	Male	Not Rearrested
21992	prisoner21992	Male	Rearrested
552	prisoner00552	Male	Rearrested
11181	prisoner11181	Male	Rearrested
36413	prisoner36413	Male	Not Rearrested
70322	prisoner70322	Female	Not Rearrested
24223	prisoner24223	Male	Rearrested
21177	prisoner21177	Male	Rearrested

Example: Contingency Table with Counts

Table 1: Rearrests Within 1 Year of Release from Prison by Sex

sex	Rearrested	Not Rearrested	Total
Male	28756	36748	65504
Female	2785	5311	8096
Total	31541	42059	73600

Example: Count Table Code

```
df |>
  tabyl(sex, reoffend) |>
  adorn_totals(where = c('row', 'col')) |>
  kbl(full_width = F,
      caption = 'Rearrests Within 1 Year of Release from Pr
  kable_classic()
```

Example: Independence

Is a prisoner being rearrested within 1 year of their release independent of their sex?

Remember: when 2 events are independent,
$P(B) \approx P(B|A) \approx P(B|A')$

$$
\begin{aligned}
P(\text{rearrestedandmale}) &= P(\text{male}) \times P(\text{rearrested} \,|\, \text{male}) \\
&\approx P(\text{male}) \times P(\text{rearrested})
\end{aligned}
$$

Example: Contingency Table with Proportions by Column

$P(\text{male}) = 0.89$

Table 2: Rearrests Within 1 Year of Release from Prison by Sex

sex	Rearrested	Not Rearrested	Total
Male	0.912	0.874	0.89
Female	0.088	0.126	0.11
Total	1.000	1.000	1.00

Example: Proportions by Column Code

```
df |>
  tabyl(sex, reoffend) |>
  adorn_totals(where = c('row', 'col')) |>
  adorn_percentages(denominator = 'col') |>
  kbl(full_width = F, digits = 3,
      caption = 'Rearrests Within 1 Year of Release from Pr
  kable_classic()
```

Example: Contingency Table with Proportions by Row

$P(\text{rearrested}) = 0.429$

Table 3: Rearrests Within 1 Year of Release from Prison by Sex

sex	Rearrested	Not Rearrested	Total
Male	0.439	0.561	1
Female	0.344	0.656	1
Total	0.429	0.571	1

Example: Proportions by Row Code

```
df |>
  tabyl(sex, reoffend) |>
  adorn_totals(where = c('row', 'col')) |>
  adorn_percentages(denominator = 'row') |>
  kbl(full_width = F, digits = 3,
      caption = 'Rearrests Within 1 Year of Release from Pr
  kable_classic()
```

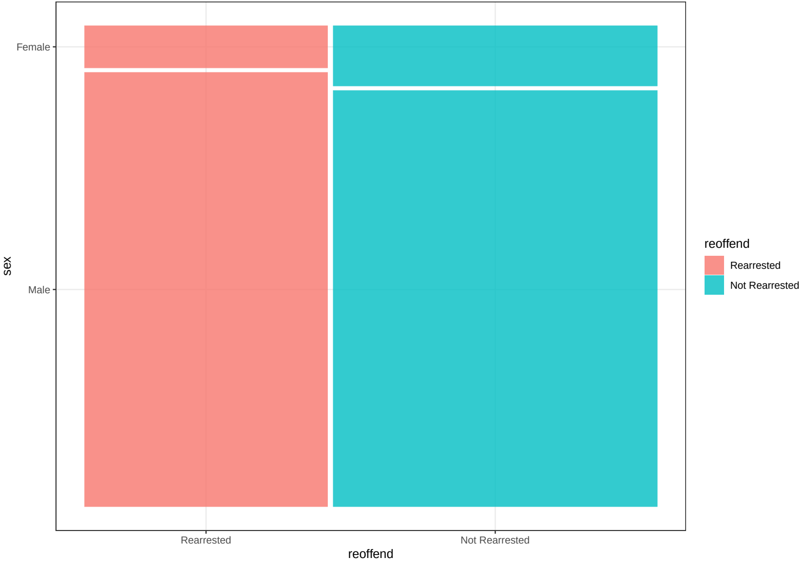
Example: Independence Calculations

Does our observed probability $P(\text{rearrestedandmale}) = 0.439$ approximate the probability under an independence model, $P(\text{rearrested}) \times P(\text{male})$?

$$\begin{aligned} P(\text{rearrested}) \times P(\text{male}) &= 0.429 \times 0.890 \\ &= 0.382 \end{aligned}$$

Does $0.439 \approx 0.382$?

Example: Mosaic Plot



Point Estimation

- We are interested in population-level parameters
- Complete populations are difficult (or impossible) to collect data from
- The ***sample statistic*** can be used as a ***point estimate*** for a population parameter in sufficiently large samples

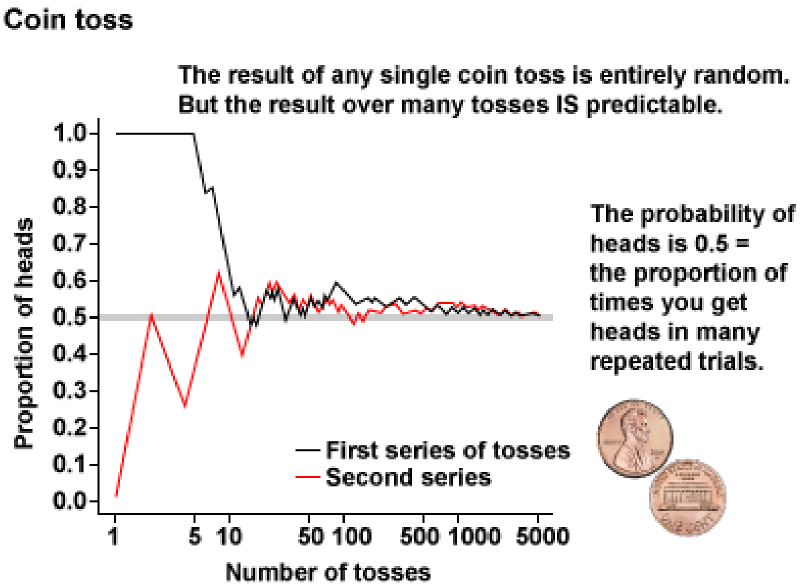
Population Parameters versus Sample Statistics

Measure	Parameter	Statistic
Mean	μ	$\bar{x}$
Proportion	p	$\hat{p}$
Difference in Means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Difference in Proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$
Standard Deviation	σ	s
Correlation	ρ	r

Sampling Distribution

- A ***sampling distribution*** is a distribution of sample statistics for different samples of the same size from the same population
- Not actually observed in the real world, a "hypothetical" population distribution
- Sampling distributions also have location and scales

Example: Law of Large Numbers



Distribution of Sample Proportions

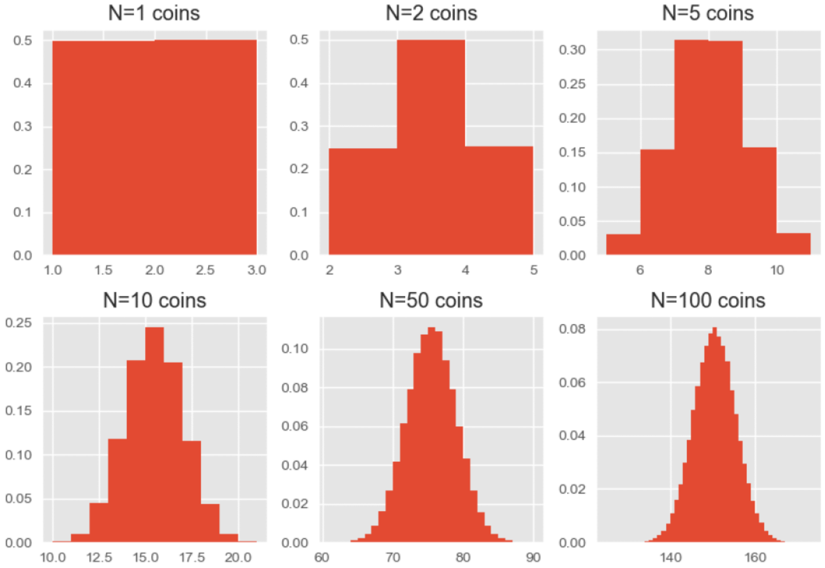


Figure 3: Sampling distribution of sample proportions of coin flips by sample size.

Uncertainty & Sampling Error

- What are sources of error we COULD know?
- What are sources of error we CAN'T know?

Central Limit Theorem

- A distribution of sample means approximates a normal distribution as the sample size gets larger
- Requires ***independent, identically distributed (I.I.D)*** variables
- Each observation is *independent* of the next
- Each observation comes from the same source distribution
- Sample size must be sufficient for estimates to be valid

Standard Error of the Sample Statistic

- ***Standard error (SE)*** is the *standard deviation* of the sample statistic
- Describes the scale (i.e. variability, sampling error) of the sampling distribution
- For a distribution of sample means, $SE = \frac{\sigma}{\sqrt{n}}$
- For a distribution of sample proportions, $SE = \sqrt{\frac{p(1-p)}{n}}$

As n increases, the standard error SE decreases.

What do sample means have to do with proportions?

- For a binary category, if you coded the 2 classes as 0 and 1, the sample mean $\bar{x}$ equals the sample proportion $\hat{p}$ of 1's.
- Example: You flip a coin 5 times, and it's tails 3 times.

$$\hat{p} = \frac{\text{count(tails)}}{\text{count(flips)}} = \frac{3}{5}$$

$$\bar{x} = \frac{\text{sum}(0, 0, 1, 1, 1)}{n} = \frac{3}{5}$$

Central Limit Theorem for Sample Proportions

- ***Assumption***: Sample proportions will be nearly normally distributed with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ (i.e. standard error)

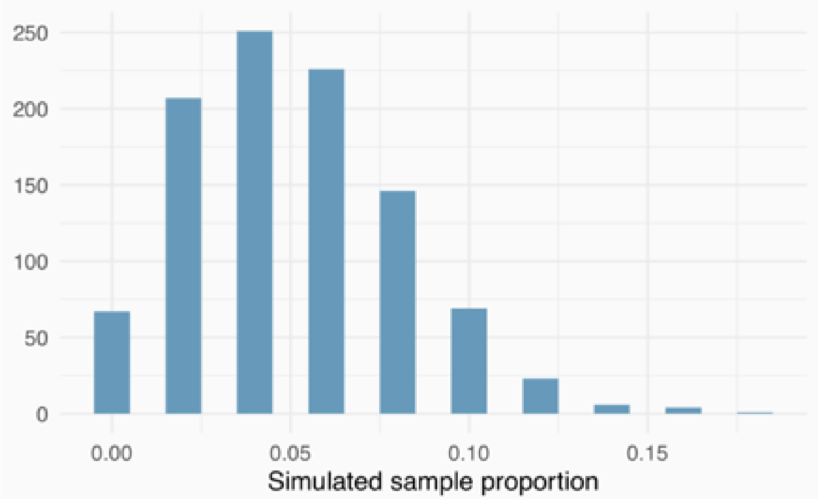
$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Requirements for CLT for Proportions

- Need a sufficiently large sample size!
- $np > 10$
- $n(1-p) > 10$
- Need ***i.i.d.*** samples

Example: Small Sample Size

In this population, $p = 0.05$ and we take random samples of size $n = 50$. Will the distribution of sample proportions be nearly normal?



Sample Size versus Proportion

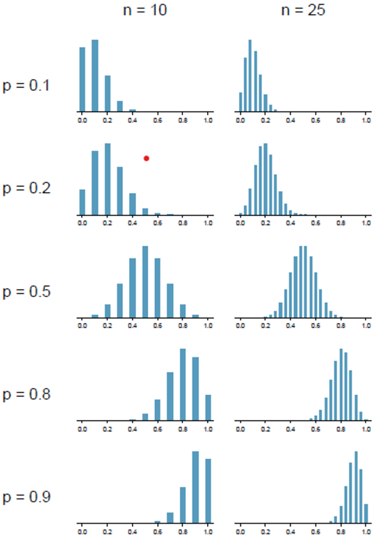


Figure 5.4: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 10$ and $n = 25$.

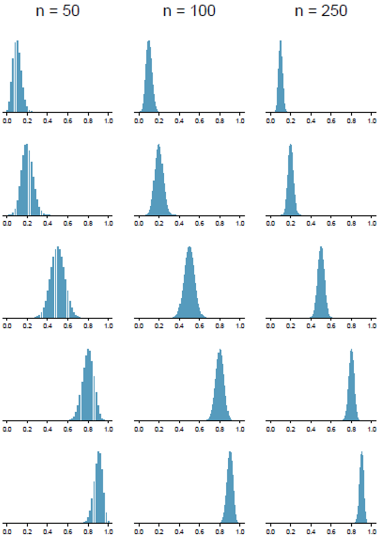


Figure 5.5: Sampling distributions for several scenarios of p and n.
Rows: $p = 0.10$, $p = 0.20$, $p = 0.50$, $p = 0.80$, and $p = 0.90$.
Columns: $n = 50$, $n = 100$, and $n = 250$.

Central Limit Theorem for Sample Means

- ***Assumption***: Sample means will be nearly normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (standard error)

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Requirements for CLT for Means

- Need a sufficiently large sample size!
- $n > 30$ when underlying distribution is normal
- Need larger n when underlying distribution is skewed
- Need ***i.i.d.*** samples

Example: Small Sample Size

If our population has the distribution $N(30, 10)$ and we take random samples of size $n = 20$. Will the distribution of sample means be nearly normal with the ***sampling distribution*** $N(30, \frac{10}{\sqrt{20}})$?

Statistical Inference and Hypothesis Testing

- We use sample statistics to describe study populations and estimate parameters using sampling distributions
- We also describe the variability of our measure and quantify our uncertainty regarding our estimate
- We use the overlap between theoretical distributions to decide how meaningful the differences between groups are

Example: COVID-19 Vaccination & Myocarditis

- Study in Denmark of 4,931,775 individuals ages 12+ or older from 2020-10-01 to 2021-10-05
- 3,482,295 individuals in the study were vaccinated and 269 developed myocarditis. 48 individuals were vaccinated *and* developed myocarditis.
- Does the probability of developing myocarditis vary by whether or not someone has been vaccinated against COVID-19?

Example: Estimating the Population Probability of Myocarditis

If 269 of the 4,931,775 individuals sampled regardless of vaccination status, what is our ***point estimate*** (i.e. sample statistic) for the probability of developing myocarditis in our study population?

$$
\begin{aligned}
\hat{p} &= \frac{\text{count(cases)}}{\text{count(subjects)}} \\
&= \frac{269}{4,931,775} \\
&= 0.00005
\end{aligned}
$$

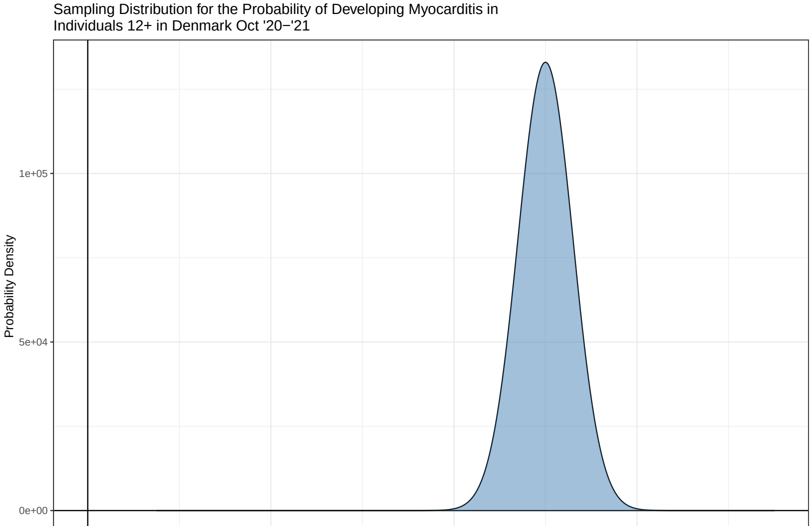
Example: Standard Error of the Measurement

If 269 of the 4,931,775 individuals sampled regardless of vaccination status, what is the ***standard error*** of our sample statistic $\hat{p} = 0.00005$, the probability of developing myocarditis in our study population?

$$
\begin{aligned}
SE_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \\
&= \sqrt{\frac{0.00005(1-0.00005)}{4,931,775}} \\
&= 0.000003
\end{aligned}
$$

Example: Myocarditis (All) Sampling Distribution

The proportion of all individuals in the study population who develop myocarditis has a sampling distribution of $N(0.00005, 0.000003)$.



Example: 68-95-99.7 Rule

If my population has the theoretical sampling distribution $p \sim N(0.00005, 0.000003)$, in what range do 95% of the theoretical sample means occur?

95% Confidence Interval $= (0.000044, 0.000056)$

95% of the theoretical sample means in this sampling distribution are between 0.000044 and 0.000056.

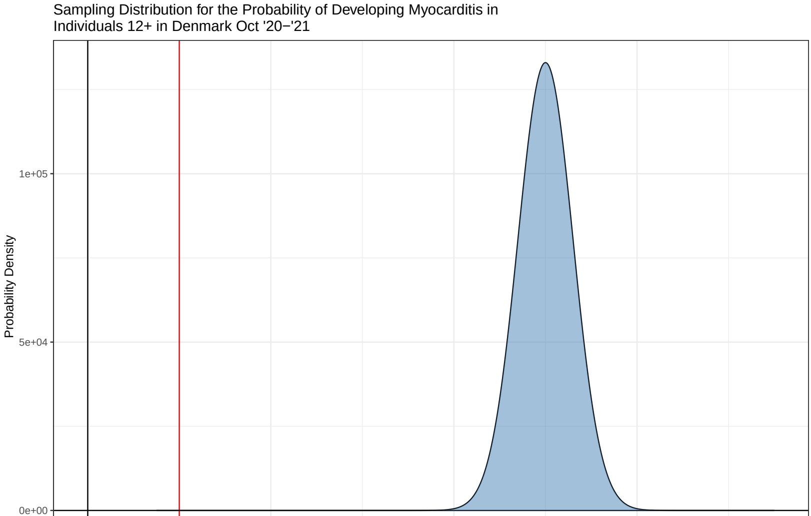
Example: Probability of Myocarditis in Vaccinated

If 48 of the 3,482,295 vaccinated individuals developed myocarditis, what is our ***point estimate*** (i.e. sample statistic) for the probability of developing myocarditis in our vaccinated population?

$$\begin{aligned}
\hat{p} &= \frac{\text{count}(\text{vaccinatedcases})}{\text{count}(\text{vaccinatedsubjects})} \\
&= \frac{48}{3,482,295} \\
&= 0.00001
\end{aligned}$$

Example: Vaccinated Myocarditis Proportion vs Population Proportion

How unusual is this sample statistic given our population parameter?



Example: Sample Proportion Percentile

What's the probability of observing a sample proportion as small or smaller than 0.00001?

```
pnorm(0.00001, mean = 0.00005, sd = 0.000003)
```

```
[1] 7.406413e-41
```

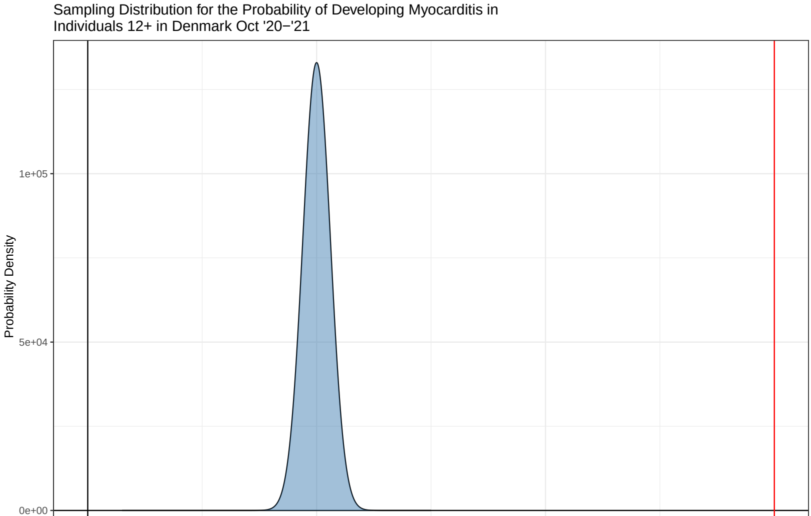
Example: Probability of Myocarditis in Unvaccinated

If 221 of the 1,449,480 unvaccinated individuals developed myocarditis, what is our ***point estimate*** (i.e. sample statistic) for the probability of developing myocarditis in our unvaccinated population?

$$\begin{aligned}
\hat{p} &= \frac{\text{count}(\text{unvaccinatedcases})}{\text{count}(\text{unvaccinatedsubjects})} \\
&= \frac{221}{1,449,480} \\
&= 0.00015
\end{aligned}$$

Example: Vaccinated Myocarditis Proportion vs Population Proportion

How unusual is this sample statistic given our population parameter?



Example: Sample Proportion Percentile

What's the probability of observing a sample proportion greater than 0.00015?

```
pnorm(0.00015, mean = 0.00005, sd = 0.000003,
      lower.tail = F)
```

```
[1] 6.352273e-244
```