Class 17

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-12

Recap: The Central Limit Theorem

- A distribution of multiple sample means approximates a normal distribution as the sample size for each mean gets larger

Recap: The Central Limit Theorem

- A distribution of multiple sample means approximates a normal distribution as the sample size for each mean gets larger
- If you take an infinite number of samples of size n from a population, the *sample statistics* (i.e. means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_\infty$) have a probability distribution (i.e. the ***sampling distribution***) that is about normal

Recap: CLT Requirements

- Requires at least 10 success/failures each for $\hat{p}$

Recap: CLT Requirements

- Requires at least 10 success/failures each for $\hat{p}$
- Requires at least $n > 30$ for $\bar{x}$

Recap: CLT Requirements

- Requires at least 10 success/failures each for $\hat{p}$
- Requires at least $n > 30$ for $\bar{x}$
- Requires independent observations

Recap: CLT Requirements

- Requires at least 10 success/failures each for $\hat{p}$
- Requires at least $n > 30$ for $\bar{x}$
- Requires independent observations
- Requires identically distributed (i.i.d.) observations

Recap: The CLT in Practice

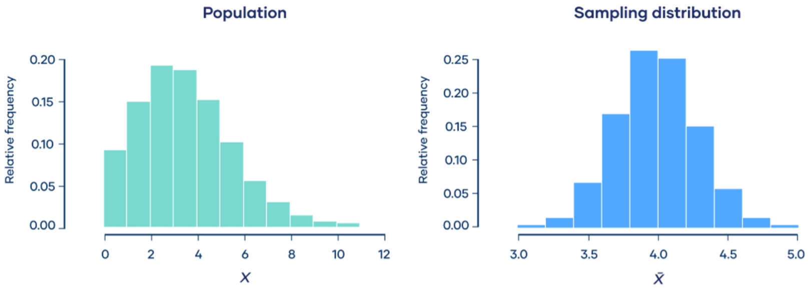


Figure 1: The ***sampling distribution*** of an infinite number of ***sample statistics*** from a population approximates a normal distribution.

Recap: Standard Error of the Sample Statistic

- ***Standard error (SE)*** is the *standard deviation* of the *sample statistic* in a theoretical *sampling distribution*
- If you took an infinite number of samples from a known distribution, the ***standard error*** is the standard deviation of the means of those samples
- Describes the scale (i.e. variability, sampling error) of the sampling distribution

Recap: Calculating the standard error

- For a distribution of sample means, $SE = \frac{\sigma}{\sqrt{n}}$

Recap: Calculating the standard error

- For a distribution of sample means, $SE = \frac{\sigma}{\sqrt{n}}$
- For a distribution of sample proportions, $SE = \sqrt{\frac{p(1-p)}{n}}$

Recap: Calculating the standard error

- 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.

Recap: Calculating a Z-Score

A ***Z-score*** indicates how many standard deviations σ away from the mean μ a given observation is.

$$\begin{aligned} Z &= \frac{\text{observed value} - \text{mean}}{\text{standard deviation}} \\ &= \frac{x - \mu}{\sigma} \end{aligned}$$

Recap: Accuracy vs Precision

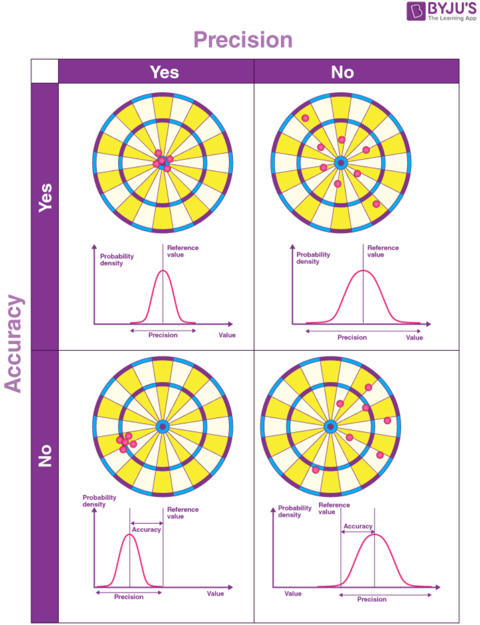
Recap: Accuracy vs Precision

- Accuracy describes how similar an observation or statistic is to the “true” population parameter

Recap: Accuracy vs Precision

- Accuracy describes how similar an observation or statistic is to the "true" population parameter
- Precision describes how similar the observations or statistics in a distribution are to each other (i.e. the variability of the estimates)

Recap: Accuracy & Precision of Estimates



Recap: Point Estimates & Confidence Intervals

- A ***point estimate*** describes the ***location*** of an estimate or distribution

Recap: Point Estimates & Confidence Intervals

- A ***point estimate*** describes the ***location*** of an estimate or distribution
- A ***confidence interval*** describes the ***scale*** of an estimate or distribution

Recap: Point Estimates & Confidence Intervals

- A ***point estimate*** describes the ***location*** of an estimate or distribution
- A ***confidence interval*** describes the ***scale*** of an estimate or distribution
- The ***confidence threshold*** or ***confidence level*** describes our uncertainty regarding these values

Recap: Confidence Intervals

A ***confidence interval*** is a numerical range *inside* which a statistic is expected to occur with a given probability $1 - \alpha$ (alpha) in any theoretical sample from a given population

Recap: Confidence Intervals

A ***confidence interval*** is a numerical range *inside* which a statistic is expected to occur with a given probability $1 - \alpha$ (alpha) in any theoretical sample from a given population

- $1 - \alpha$ is the ***confidence level*** and is often expressed as a %

Recap: Confidence Intervals

A ***confidence interval*** is a numerical range *inside* which a statistic is expected to occur with a given probability $1 - \alpha$ (alpha) in any theoretical sample from a given population

- $1 - \alpha$ is the ***confidence level*** and is often expressed as a %
- ***This is only true if your assumptions about the population hold.***

Recap: Confidence Intervals in Practice

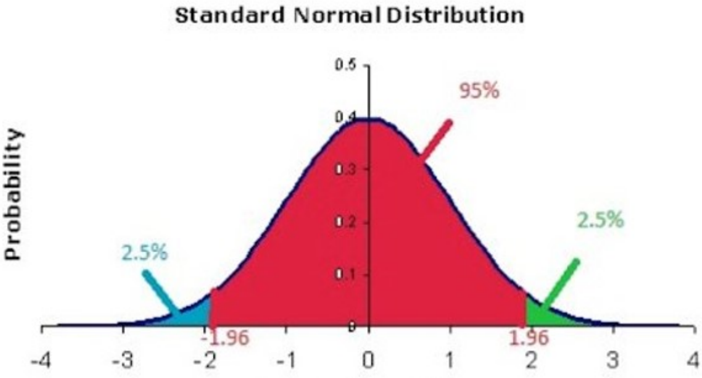


Figure 3: Properties of known distributions, like the 68-95-99.7 Rule, are used to calculate the bounds of a confidence interval.

Recap: Confidence Intervals & Z^*

- A confidence interval is defined as $\text{point estimate} \pm \text{margin of error}$
- $\text{margin of error} = Z^* \times SE$
- $Z^* = \text{Z-Score}_{\alpha/2}$

Recap: Assumptions

1. The random process follows a known distribution which we can use to model the process and draw inferences.

Recap: Assumptions

1. The random process follows a known distribution which we can use to model the process and draw inferences.
2. Your data is ***reliable***, so your sample statistics are ***reliable*** estimations of your sample population distribution.

Recap: Assumptions

1. The random process follows a known distribution which we can use to model the process and draw inferences.
2. Your data is ***reliable***, so your sample statistics are ***reliable*** estimations of your sample population distribution.
3. Your data is ***valid***, so a sampling distribution based on your sample statistics is a ***valid*** estimation of the "true" distribution in the study population.

Recap: Assumptions

1. The random process follows a known distribution which we can use to model the process and draw inferences.
2. Your data is ***reliable***, so your sample statistics are ***reliable*** estimations of your sample population distribution.
3. Your data is ***valid***, so a sampling distribution based on your sample statistics is a ***valid*** estimation of the "true" distribution in the study population.
4. Your data is ***generalizable***, so your estimated sampling distribution for your study population is ***generalizable*** as the "true" sampling distribution for your target population

Statistical Inference and Hypothesis Testing

- We use sample statistics to describe sample populations and estimate the parameters of the study population's sampling distribution

Statistical Inference and Hypothesis Testing

- We use sample statistics to describe sample populations and estimate the parameters of the study population's sampling distribution
- We also describe the variability of our measure and quantify our uncertainty regarding our estimate

Statistical Inference and Hypothesis Testing

- We use sample statistics to describe sample populations and estimate the parameters of the study population's sampling distribution
- 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

Hypothesis Testing Framework

- $\mathbf{H_0}$: The "Null" Hypothesis
  - Represents a position of skepticism, *nothing* is happening here
  - "There is *not* an association between process A and B"

Hypothesis Testing Framework

- $\mathbf{H_0}$: The "Null" Hypothesis
  - Represents a position of skepticism, *nothing* is happening here
  - "There is *not* an association between process A and B"
- $\mathbf{H_A}$: The "Alternative" Hypothesis
  - The complement of H_0, *something* is happening here
  - "There *is* an association between process A and B"

Conducting a hypothesis test

- Begin by *assuming* H_0 is the "true" state

Conducting a hypothesis test

- Begin by *assuming* H_0 is the "true" state
- Calculate *the probability that you would see results as extreme or more extreme* than what you saw in your study, assuming the distribution under H_0

Conducting a hypothesis test

- Begin by *assuming* H_0 is the “true” state
- Calculate *the probability that you would see results as extreme or more extreme* than what you saw in your study, assuming the distribution under H_0
- The lower the probability, the less likely it is that we would see these results if H_0 was the “true” state of our population

Conducting a hypothesis test

- Begin by *assuming* H_0 is the "true" state
- Calculate *the probability that you would see results as extreme or more extreme* than what you saw in your study, assuming the distribution under H_0
- The lower the probability, the less likely it is that we would see these results if H_0 was the "true" state of our population
- If the probability is sufficiently low, we *reject* $\mathbf{H_0}$ and *accept* $\mathbf{H_A}$

Significance Level/Threshold

- α is also called the ***significance level***

Significance Level/Threshold

- α is also called the ***significance level***
- The probability below which you will reject the null hypothesis

Significance Level/Threshold

- α is also called the ***significance level***
- The probability below which you will reject the null hypothesis
- Predetermined before doing hypothesis test (often $p < 0.05$)

Significance Level/Threshold

- α is also called the ***significance level***
- The probability below which you will reject the null hypothesis
- Predetermined before doing hypothesis test (often $p < 0.05$)
- Also the probability of rejecting the null hypothesis when H_0 is true (i.e. ***Type I Error*** or ***false positive rate***)

Decision Errors

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Inference for a Single Proportion

Central Limit Theorem for Proportions: sample proportions $\hat{p}$ will be nearly normally distributed with the mean equal to the population proportion ($\mu = p$) and the standard deviation equal to the standard error for a proportion ($\sigma = \sqrt{\frac{p(1-p)}{n}}$), such that $\hat{p} \sim N(\mu = p, \sigma = SE_p)$).

Inference for a Single Proportion

Central Limit Theorem for Proportions: sample proportions $\hat{p}$ will be nearly normally distributed with the mean equal to the population proportion ($\mu = p$) and the standard deviation equal to the standard error for a proportion ($\sigma = \sqrt{\frac{p(1-p)}{n}}$), such that $\hat{p} \sim N(\mu = p, \sigma = SE_p)$).

Assumptions: independence, identically distributed, 10+ successes/failures each

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

Research Question: Is a hurricane more likely to hit the continental US in 2024?

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

What is the study population?

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

What is the study population?

All hurricanes which formed in the Atlantic Ocean with the potential to make landfall in the continental US, for which we have records.

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

What is the sample population?

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

What is the sample population?

298 hurricanes which formed in the Atlantic Ocean between 1980-2023 with the potential to make landfall in the continental United States.

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

What is the target population?

Example

From 1980-2023, 709 tropical cyclones have formed in the Atlantic Ocean. 298 of those tropical cyclones developed into hurricanes, and 72 of those hurricanes made landfall in the continental US.

So far in 2024, 13 tropical cyclones have formed in the Atlantic Ocean. 9 of those tropical cyclones developed into hurricanes, and 2 of those hurricanes made landfall in the continental US.

What is the target population?

Future hurricanes which form in the Atlantic Ocean with the potential to make landfall in the continental US.

Example

Is it reasonable to assume that the sample statistics from the data will reliably describe the observed distribution in the sample population?

Example

Is it reasonable to assume that the sample statistics from the data will reliably describe the observed distribution in the sample population?

Is it reasonable to assume that the sample statistics will be a valid estimation of the sampling distribution in the study population?

Example

Is it reasonable to assume that the sample statistics from the data will reliably describe the observed distribution in the sample population?

Is it reasonable to assume that the sample statistics will be a valid estimation of the sampling distribution in the study population?

Is it reasonable to assume that the estimated sampling distribution for the study population will be generalizable to the unobserved distribution in the target population?

Example

Is it reasonable to assume that the population parameters can be modeled using a normal distribution?

Example

Is it reasonable to assume that the population parameters can be modeled using a normal distribution?

- Are the observations *independent*?

Example

Is it reasonable to assume that the population parameters can be modeled using a normal distribution?

- Are the observations *independent*?
- Are the observations *identically distributed*?

Example

Is it reasonable to assume that the population parameters can be modeled using a normal distribution?

- Are the observations *independent*?
- Are the observations *identically distributed*?
- Is the *sample size sufficient*?

Example

If we assume our data is reliable, then our sample statistics will be accurate estimations of the underlying distribution in the sample population.

Example

If we assume our data is reliable, then our sample statistics will be accurate estimations of the underlying distribution in the sample population.

If we assume our data is valid, then we can use our sample statistics to *infer* the sampling distribution for the study population.

Example

If we assume our data is reliable, then our sample statistics will be accurate estimations of the underlying distribution in the sample population.

If we assume our data is valid, then we can use our sample statistics to *infer* the sampling distribution for the study population.

If we assume our data is generalizeable, then we can use our sampling distribution to *test the hypothesis* in the target population.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 1: Calculate the sample statistic.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 1: Calculate the sample statistic.

$$\hat{p} = \frac{72}{298} = 0.242$$

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 2: Estimate the sampling distribution.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 2: Estimate the sampling distribution.

$$SE = \sqrt{\frac{0.242(1-0.242)}{298}} = 0.025$$

The sampling distribution for $\hat{p}$ approximates the normal distribution $N(24.2, 2.5)$.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 3: Calculate Z^* for the confidence threshold $\alpha = 0.05$.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 3: Calculate Z^* for the confidence threshold $\alpha = 0.05$.

$$Z^* = Z_{\alpha/2}$$

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 3: Calculate Z^* for the confidence threshold $\alpha = 0.05$.

$$Z^* = Z_{\alpha/2}$$

```
qnorm(0.05 / 2)
```

```
[1] -1.959964
```

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 4: Construct a 95% confidence interval.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 4: Construct a 95% confidence interval.

$\text{point estimate} \pm Z^* \times SE$

```r
24.2 - qnorm(0.05 / 2) * 2.5
```

```
[1] 29.09991
```

```r
24.2 + qnorm(0.05 / 2) * 2.5
```

```
[1] 19.30009
```

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 4: Construct a 95% confidence interval.

$\text{point estimate} \pm Z^* \times SE$

```
24.2 - qnorm(0.05 / 2) * 2.5
```

```
[1] 29.09991
```

```
24.2 + qnorm(0.05 / 2) * 2.5
```

```
[1] 19.30009
```

With 95% confidence, the probability of a hurricane making landfall in the continental US is 19.3% to 29.1%.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 5: Assume the null hypothesis.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 5: Assume the null hypothesis.

H_0: The probability of a hurricane making landfall in 2024 is 24.2% ($\hat{p} = 24.2\%$).

H_A The probability of a hurricane making landfall in 2024 is *not* 24.2% ($\hat{p} \neq 24.2\%$).

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 6: Calculate the sample statistic.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 6: Calculate the sample statistic.

$$\begin{aligned}\hat{p} &= \frac{2}{9} \\ &= 0.222\end{aligned}$$

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 7: Calculate the test statistic under H_0.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 7: Calculate the test statistic under H_0.

$$
\begin{aligned}
Z &= \frac{\hat{p} - p}{SE} \\
&= \frac{22.2 - 24.2}{2.5} \\
&= -0.8
\end{aligned}
$$

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 8: Calculate the p-value under H_0.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 8: Calculate the p-value under H_0.

```
pnorm(-0.8, mean = 0, sd = 1) * 2
```

```
[1] 0.4237108
```

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 9: Reject or fail to reject the null hypothesis.

Example

Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 9: Reject or fail to reject the null hypothesis.

The p-value for the observed data under the null hypothesis is $p = 0.423$. As $p > \alpha$ ($\alpha = 0.05$), this is *not* sufficient evidence of a difference.

Example

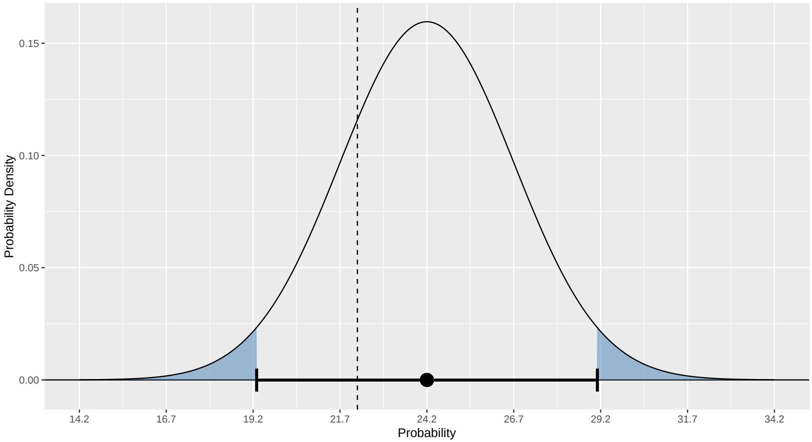
Based on the data from 1980-2023, what is the average probability that a hurricane makes landfall in the continental US?

Step 9: Reject or fail to reject the null hypothesis.

The p-value for the observed data under the null hypothesis is $p = 0.423$. As $p > \alpha$ ($\alpha = 0.05$), this is *not* sufficient evidence of a difference.

We fail to reject the null hypothesis that the probability of a hurricane making landfall in 2024 is 24.2%.

Example



0.15
0.10
0.05
0.00
Probability Density
14.2
16.7
19.2
21.7
24.2
26.7
29.2
31.7
34.2
Probability