

Z-tests

Grinnell College

November 2025

Sampling Distributions for...

- ▶ Numeric data with known variance
 - ▶ Z-test
 - ▶ Confidence intervals
- ▶ Numeric data with unknown variance
 - ▶ t-test
 - ▶ Confidence intervals
- ▶ Difference of two numeric variables (with unknown variance) and ANOVA
 - ▶ t-tests and ANOVA
 - ▶ Confidence intervals

Next Unit

- ▶ Categorical data with proportions
 - ▶ 1 proportion (test + confidence interval)
 - ▶ 2 proportions (test + confidence intervals; super brief to say I actually did show it to you)
 - ▶ χ^2 -test for tables
- ▶ Regression
 - ▶ t-tests and conf. int. for coefficients
 - ▶ Multiple Regression
- ▶ Logistic Regression

Starting Point

We talked about how statistics calculated from a random sample will have a long term behavior known as a sampling distribution.

The sample mean \bar{x} actually has a distribution, it's...

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad (1)$$

- ▶ only true if the assumptions hold
- ▶ μ is the population mean
- ▶ σ^2 is the population variance
 - ▶ For this slide deck we will act like we know this
- ▶ n is our sample size

Assumptions

First we have to check if the assumptions are correct....

- ▶ The observations are randomly sampled
- ▶ The observations are independent and come from the same population
 - ▶ “independent and identically distributed” = “iid”
- ▶ The population is normally distributed or the sample size is large
 - ▶ How can we tell if the population is normally distributed?

Assumptions

First we have to check if the assumptions are correct....

- ▶ The observations are independent and come from the same population
 - ▶ “independent and identically distributed”
- ▶ The population is normally distributed OR the sample size is large
 - ▶ How can we tell if the population is normally distributed? We use the sample distribution which should look like the population distribution
 - ▶ The sample size being large enough kicks in the central limit theorem (next slide)

If both main bullet points are checked off we can use the normal distribution to approximate the sampling distribution

Central Limit Theorem

The **central limit theorem** (CLT) says the the long term behavior of an average will be normally distributed if the sample size (eg people in a single study) is sufficiently large

- ▶ Large sample size needed (called an asymptotic result)
- ▶ Makes the sampling distribution normal
- ▶ There is a requirement that the population distribution actually **has** a mean and variance
 - ▶ Never(?) questioned in practice
 - ▶ Take much higher level math classes for more details
 - ▶ Eg Cauchy distribution doesn't have a mean

Central Limit Theorem

What is a large sample size?

- ▶ 30 is the classic response (half joke anymore?)
- ▶ Depends on how not-normal the population distribution is
 - ▶ Again, we use the sample distribution in practice
 - ▶ If symmetric without outliers $N \approx 20$
 - ▶ Skew or some outliers $N \approx 50$
 - ▶ Really really weirdly shaped $N \approx 100$
- ▶ I personally have great faith in the CLT

Once you hit a sample size of 100 or more the CLT will do well for pretty much anything

Example 1

Over the course of a few years in the mid-2000's a collection of body measurements were taken on colonies of penguins, randomly selected, on some islands off of Antarctica. Find the sampling distribution for mean flipper length



Example 1: Assumptions

Are the observations random?



Are the observations independent?

Is the population normally distributed or is the sample size large?

Example 1: Assumptions

Are the observations random?

- ▶ Yes, it says so

Are the observations independent?

- ▶ ?

Is the population normally distributed or is the sample size large?

Ex: Penguin Assumptions

Are the observations random?

- ▶ Yes, it says so

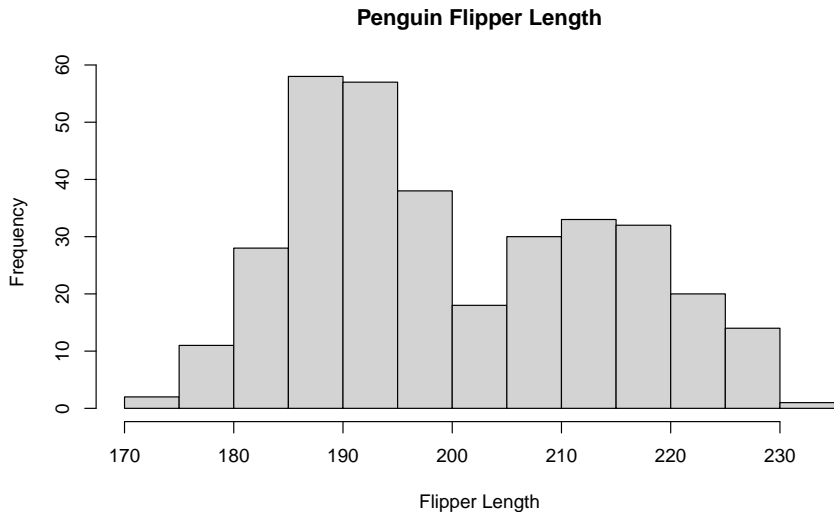
Are the observations independent?

- ▶ Maybe? There'll be genetic relatives in each colony
- ▶ And each colony could be different with different means

Is the population normally distributed or is the sample size large?

- ▶ The population.....?

Ex: Penguin Plot



Ex: Penguin Assumptions

Are the observations random?

- ▶ Yes, it says so

Are the observations independent?

- ▶ Maybe? There'll be genetic relatives in each colony
- ▶ And each colony could be different with different means

Is the population normally distributed or is the sample size large?

- ▶ The population is not normally distributed (because the sample isn't)
- ▶ The sample size is....

Ex: Penguin Assumptions

Are the observations random?

- ▶ Yes, it says so

Are the observations independent?

- ▶ Maybe? There'll be genetic relatives in each colony
- ▶ And each colony could be different with different means

Is the population normally distributed or is the sample size large?

- ▶ The population is not normally distributed (because the sample isn't)
- ▶ The sample size is....large with over 300 observations

Despite possible concerns with independence let's move forward

Sampling Distribution for a Single Mean

The **sampling distribution for a single mean** will be....

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

where...

- ▶ μ is our population's mean
 - ▶ Unknown
 - ▶ We can estimate it with our sample
 - ▶ Or we can hypothesize its value
- ▶ σ^2 is our variance
 - ▶ Population's variance
 - ▶ Almost always unknown and must be estimated
 - ▶ For this slide deck assume we know the true variance
- ▶ n is our sample size

Ex: Penguin Sampling Distribution

The resulting sampling distribution then is $N(\mu, \sigma^2/n)$ which is

$$\bar{X} \sim N(\mu, 180/344)$$

where

- ▶ μ is the mean of our population
 - ▶ We can estimate it with our sample mean (200)
 - ▶ We can hypothesize its value
- ▶ 180 is our (pop) variance
 - ▶ Assumed known before we ever started talking about penguins
 - ▶ Very unrealistic short of an ornithologist savant
- ▶ 344 is our sample size

Let's review hypothesis tests real fast and then we can try some

Z-Test

1. State our two hypothesis
2. ~~Define a thresh hold/critical value~~
3. Visualize your data (already done)
4. Check assumptions (already done)
 - ▶ Independence from the same population
 - ▶ Normal or Large N
5. State your sampling distribution
6. Test Statistic
 - ▶ Standardized value we observed
7. Probability we see that test statistic (or more extreme)
 - ▶ Calculate the probability from our (normal) sampling dist.
8. Decision is made
 - ▶ profit

Step 1: Hypothesis Statement

The **null hypothesis** is the hypothesis you are wanting to show is most likely wrong

- ▶ Denoted as H_0
- ▶ Will use $=$, \geq , or \leq

The **alternative hypothesis** is the other possible state of the world that is not the null hypothesis.

- ▶ Defined as what the null hypothesis isn't
- ▶ Will use $>$, $<$, or \neq
- ▶ Usually this is what you want to show

For means, we will use the (unknown) parameter μ

Ex: Penguin Z-test

My uncle: *The average penguin flipper length? I don't know Vinny I'd guess a foot long maybe*

Let's try to disprove my uncle's (unwilling) claim that a penguins flipper is about a foot long (305 mm).

Step 1 is to write out our hypothesis

Ex: Hypothesis Statements

You must list two hypothesis statements, the null and the alternative.

$$H_0 : \mu_{flipper} = 305$$

$$H_A : \mu_{flipper} \neq 305$$

Here we are trying to show there is a difference; that RJ's claim they the flippers are a foot long is just wrong.

Sampling Distribution's Mean for a Hypothesis Test

Critical: During a hypothesis test we assume we *know* the population mean (via H_0)

- ▶ We assume the sampling distribution is centered at the population's mean
- ▶ We make some claim about the pop's mean in H_0
- ▶ So plug in the hypothesized value from H_0 into the sampling distribution

When we judge whether the results we got are unlikely or not; we are judging in effect how believable H_0 is.

Ex: Penguin's Sampling Dist for 1 Mean HT

Initially we had...

$$\bar{X} \sim N(\mu, 180/344)$$

but we still need to replace μ with our hypothesized value so...

- ▶ $H_0 : \mu_{flipper} = 305$
- ▶ $180/334 = .54$

$$\bar{X} \sim N(305, .54)$$

Ex: Penguin's Test Statistic

For a Z-test, the test statistic is the standardized observation so....

- ▶ The sample mean was 200
- ▶ Hypothesized mean is 305
- ▶ And the standard deviation for our sampling distribution is $(\frac{\sigma^2}{n})^{1/2} = (180/334)^{1/2} = .7341$
 - ▶ This is called the **standard error**; it's the standard deviation of the sampling distribution
 - ▶ Personally I think it having it's own name is kind of.....not wise

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{200 - 305}{.7341} = 143$$

Probability we'd see this (ie p-value)

Let T be our test statistic from the previous step and Z represent a standard normal variable. Three possible situations based on the alternative hypothesis:

- ▶ $P(Z < T)$
 - ▶ $H_A: \mu < (\text{SOME NUMBER})$
 - ▶ We take the probability to the left
- ▶ $P(Z > T)$
 - ▶ $H_A: \mu > (\text{SOME NUMBER})$
 - ▶ We take the probability to the right
- ▶ $2 * P(Z > |T|)$
 - ▶ $H_A: \mu \neq (\text{SOME NUMBER})$
 - ▶ This sums the probability we'd be lower than $-|T|$ or greater than $|T|$
 - ▶ We multiply by 2 because we use symmetry to make the equation easier

Ex: Penguin's Weirdness (ie P-value)

NOTE: Because we are using \neq in our alternative hypothesis we need....

$$2 * P(Z > |T|) = 2 * P(Z > |143|)$$

Which is equivalent to...

$$P(Z < -143 \cup Z > 143) = P(Z < -143) + P(Z > 143)$$

$$2 * \text{pnorm}(143, \text{mean} = 0, \text{sd} = 1, \text{lower.tail} = \text{FALSE}) \sim 0$$

The probability that we'd see an average of 200mm given the real mean is 305mm is below R's reporting ability to calculate (eg R just says "0")