Parametric inference on means

Overview

Information on the final project

Review of parametric inference on proportions

Parametric inference on a single mean

- Distribution of statistics
- Confidence intervals for a single mean
- Hypothesis tests for a single mean

Final project

Final project: analyze your own data set

Final project report: a 5-8 page Quarto document that contains:

1. Background information:
    - What question you will address and why it is interesting
    - Where you got the data and any prior analyses
2. Descriptive plots
3. A hypothesis tests using randomization and parametric methods
4. A confidence interval using the bootstrap and parametric methods
5. A conclusion and reflection
6. Optional: an appendix with extra code
    - (appendix can go over the 8 page limit)

Example: Do beavers have the same body temperature as humans?

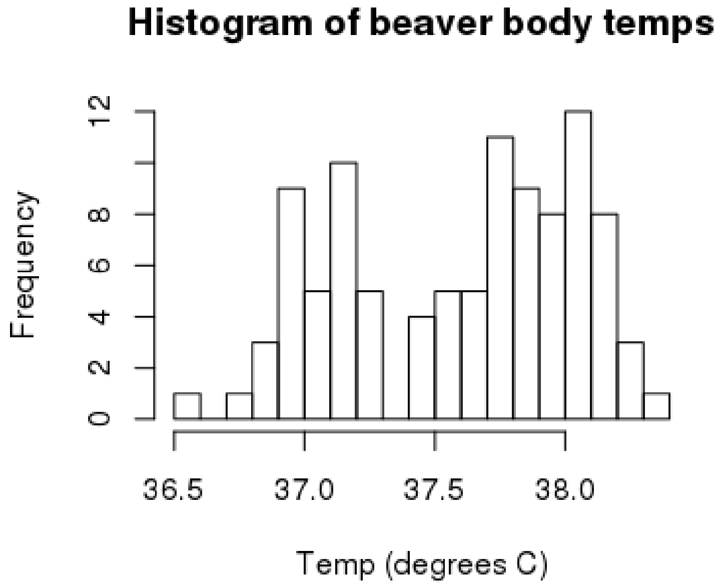
Motivation and data

Motivation: There is a labor shortage in the construction industry

- Beavers are a hard working species of animals
- If beavers have the same body temperature as humans (37℃), perhaps they can be employed in the construction industry

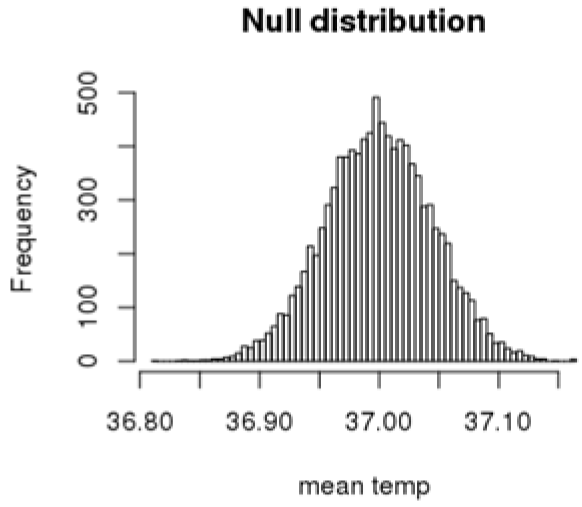
The data:

- Body temperatures collected from 400 beavers*
- Data from:
  - Lange et al (1994). In time-series analyses of beaver body temperatures. https://vincentarelbundock.github.io/Rdatasets/doc/boot/beaver.html



*not the real data

Results



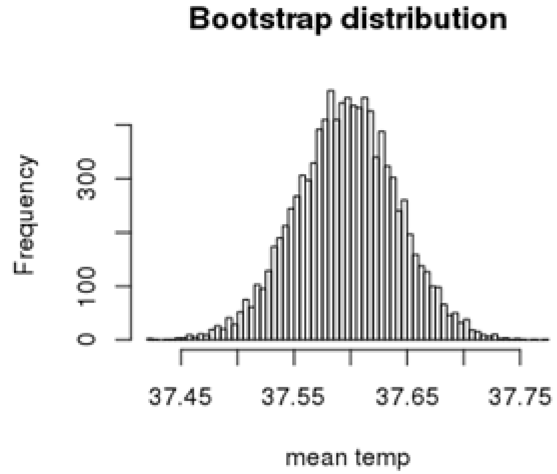
The average human body temperatures is μ = 37°C

Hypothesis test

- H_0: μ = 37 H_A: μ ≠ 37
- p-value based on a permutation test: $\overline{x}$ = 37.6,   p-value = 0
- p-value based on a t-test: t = 13.35,   df = 99,   p-value = 0

95% confidence interval for the mean beaver body temp

- Bootstrap: [37.51 37.68]
- t-distribution: [37.51 37.68]



Conclusions

Conclusion: Beavers do not seem to have the same body temperatures as humans

37℃ humans vs. 37.6℃ beavers

Implications: Due to their higher body temperatures, if beavers join the construction industry they might be too good at their jobs leading to job loss of human workers

Caveats: human body temperatures might not be exactly 37℃

Getting the final project template and loading data

A list of a few data sets you can use are on Canvas

There is a template for the final project that outlines the sections that should be in the project

You can get a final project template using the command:

```
library(SDS1000)
goto_final_project()
```

Let’s go over how to load your own data into R!

Review of inference on proportions

Central Limit Theorem for Sample Proportions

When taking samples of size n from a population with a proportion parameter π, the distribution of the sample proportions p̂ has the following characteristics:

Shape: If the sample size is sufficiently large, the distribution is reasonably normal

Center: The mean is equal to the population proportion π

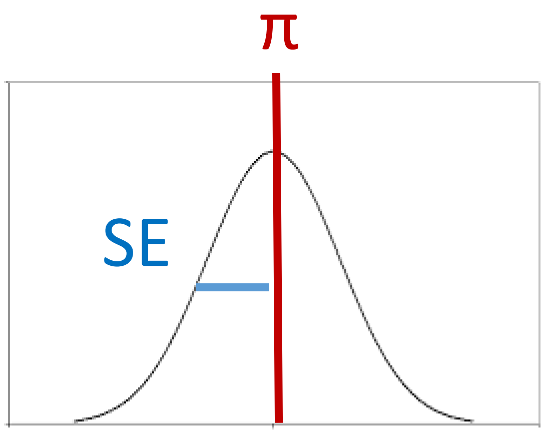
Spread: The standard error is: $SE = \sqrt{\frac{\pi(1-\pi)}{n}}$

$$\hat{P} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$$

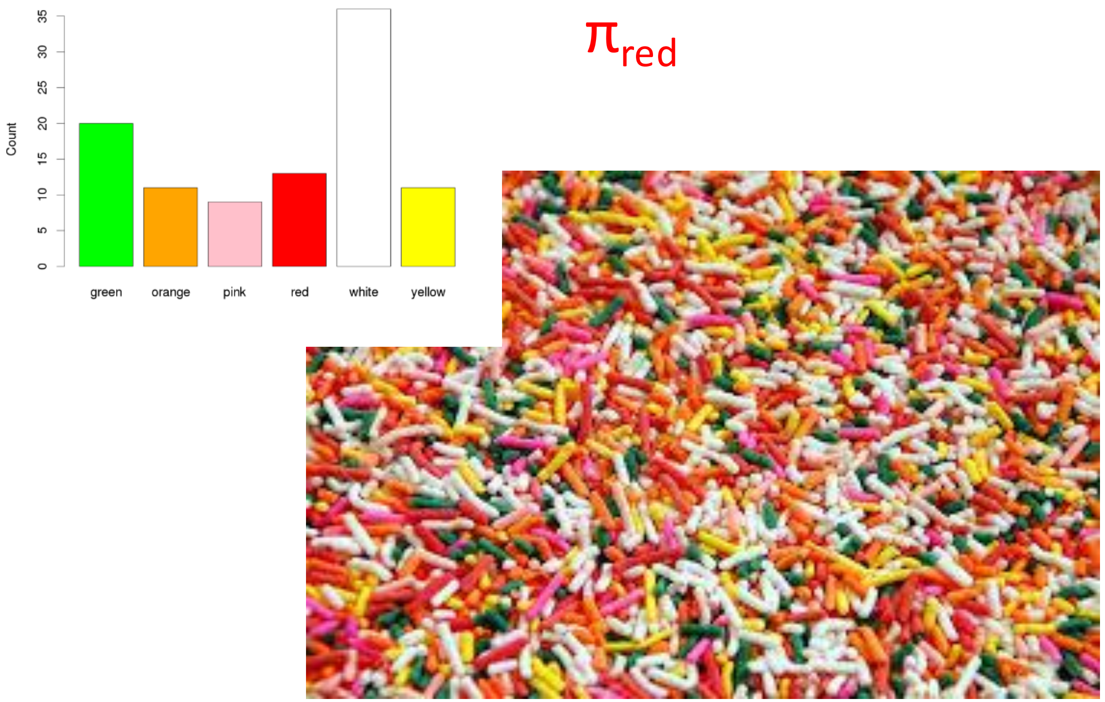
A normal distribution is a good approximation as long as:

nπ ≥ 10 and n(1 – π) ≥ 10

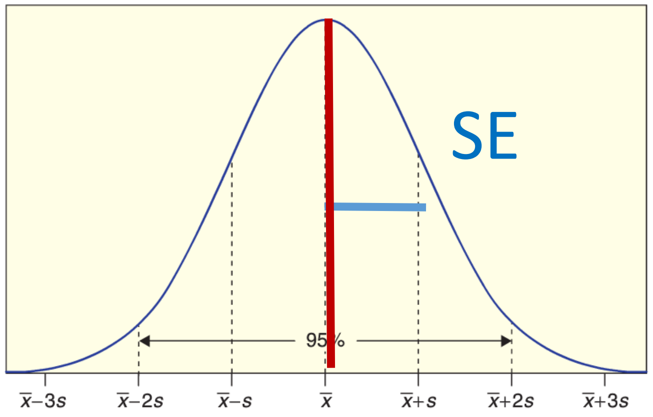
$$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$$



π_{red}



π

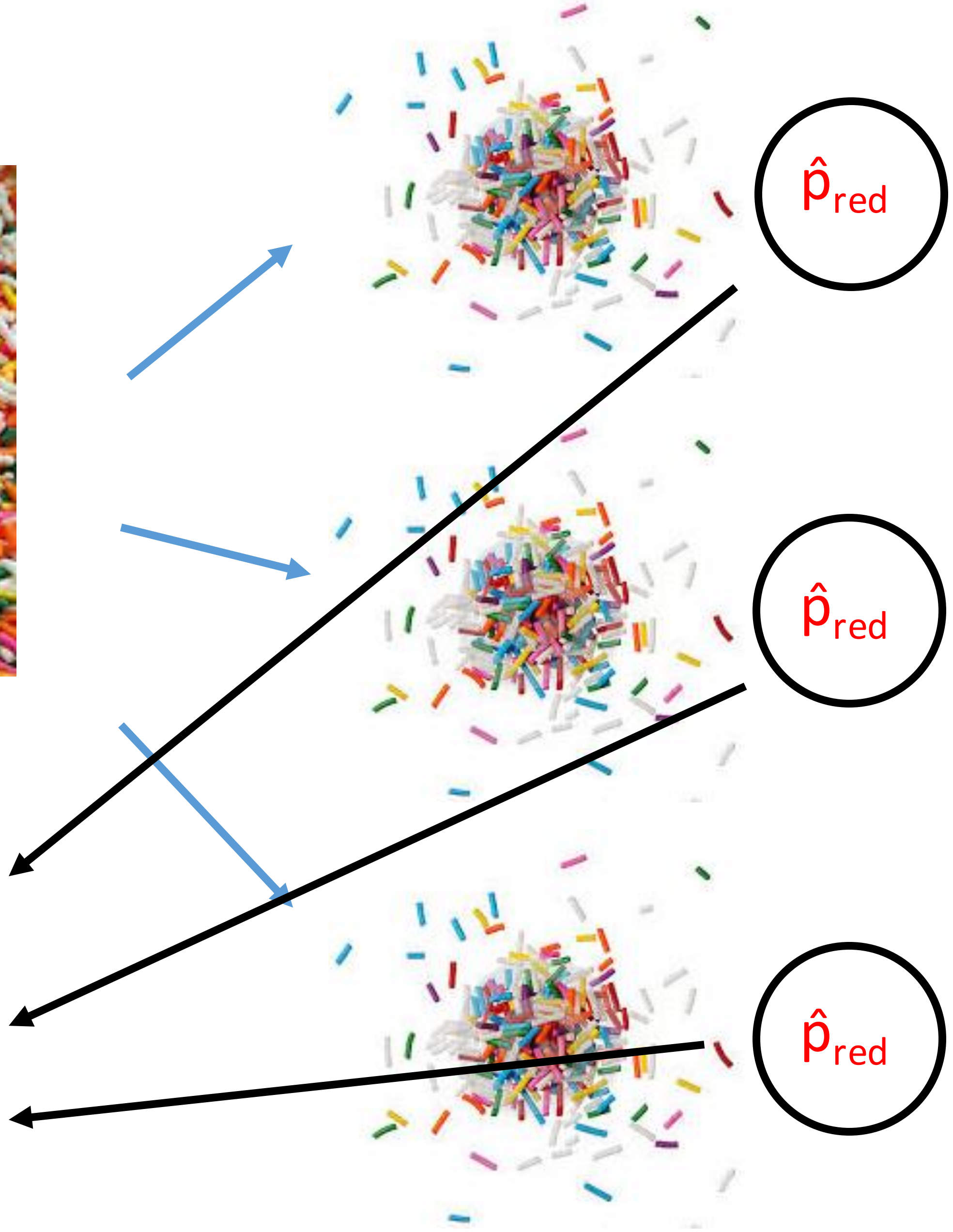


SE

Sampling distribution!

$$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$\hat{P} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$$



$\hat{p}_{red}$

$\hat{p}_{red}$

$\hat{p}_{red}$

Confidence intervals for a single proportion

Confidence interval for a single proportion

$$\hat{p} \ \pm \ z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

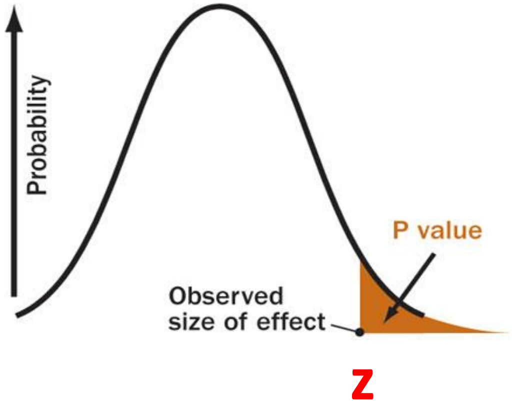
Note we are substituting p̂ for π

Hypothesis tests for a single proportion

Test statistic for a single proportion: H_0: $\pi = \pi_0$

$$z = \frac{stat_{obs} - param_0}{SE} \qquad z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Our z statistic comes from a standard normal distribution z ~ N(0, 1)



H_A: $\pi > \pi_0$

pnorm(z, 0, 1, lower.tail = FALSE)

Practice: Adult persistence of head-turning asymmetry

Background

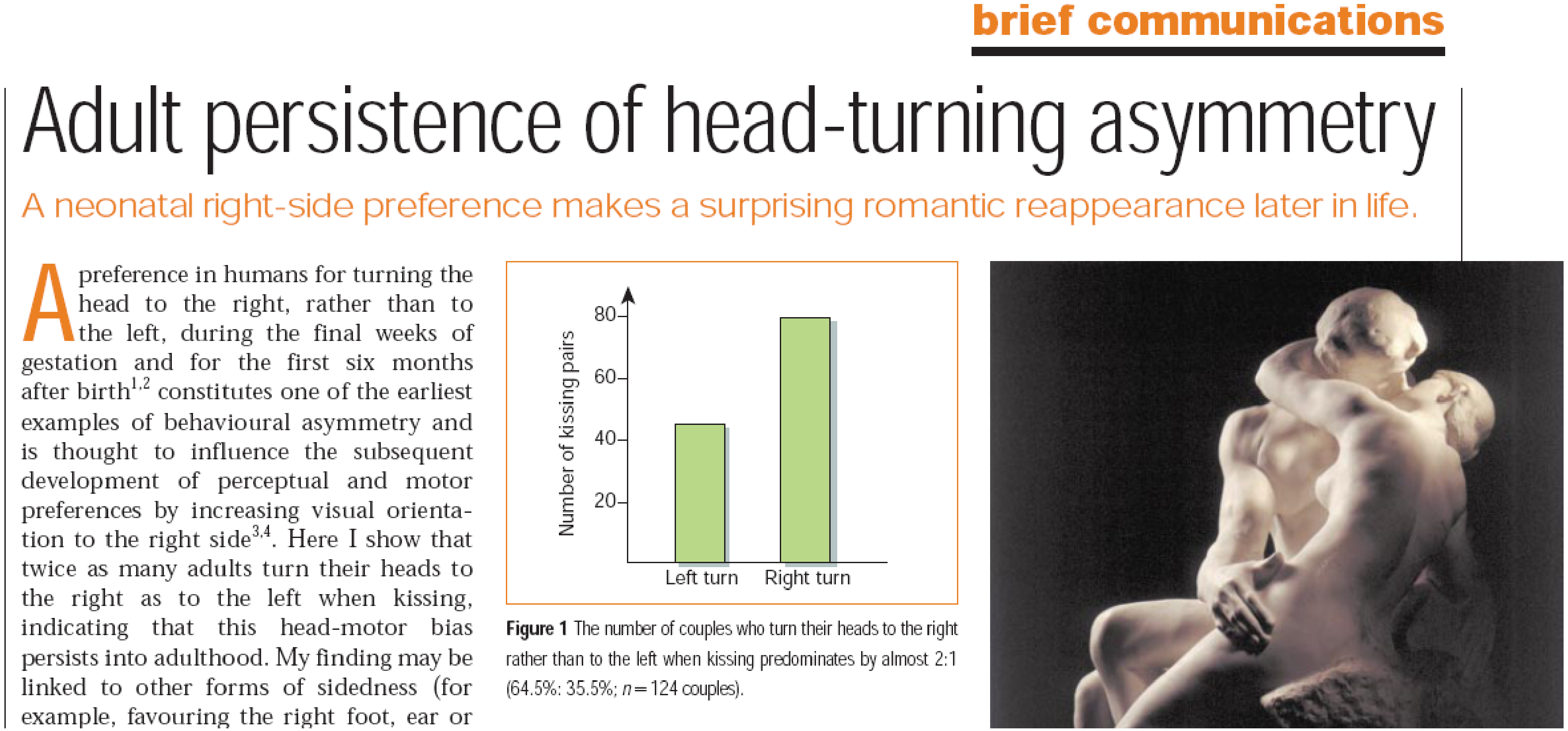
- Most people are right-handed, right eye dominant, etc.
- Biologists have suggested that human embryos tend to turn their heads to the right as well

German bio-psychologist Onur Güntürkün conjectured that this tendency manifests itself in other ways, so he studies which ways people turn their heads when they kiss

Adult persistence of head-turning asymmetry

He and his researchers observed kissing couples in public places and noted whether the couple leaned their heads to the right or left

They observed 124 couples, ages 13-70 years



brief communications

Adult persistence of head-turning asymmetry

A neonatal right-side preference makes a surprising romantic reappearance later in life.

A preference in humans for turning the head to the right, rather than to the left, during the final weeks of gestation and for the first six months after birth[1,2] constitutes one of the earliest examples of behavioural asymmetry and is thought to influence the subsequent development of perceptual and motor preferences by increasing visual orientation to the right side[3,4]. Here I show that twice as many adults turn their heads to the right as to the left when kissing, indicating that this head-motor bias persists into adulthood. My finding may be linked to other forms of sidedness (for example, favouring the right foot, ear or

Figure 1 The number of couples who turn their heads to the right rather than to the left when kissing predominates by almost 2:1 (64.5%: 35.5%; $n = 124$ couples).

Of the 124 couples observed, 80 leaned their heads to the right while kissing

- Let's run a parametric hypothesis test to assess if a higher proportion of kissers turn their heads to the right

Complete the following steps to analyze the data

1. State Null and Alternative in symbols and words

2. Calculate the observed z-statistic of interest $z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$

3-4. Use the standard normal distribution to get a p-value

5. Make a decision about whether the results are statistically significant

Let's try it in R!

Adult persistence of head-turning asymmetry

1. $H_0: \pi = 0.5$ $H_A: \pi > 0.5$ .5 * 124 = 62 > 10

2. p̂ = 80/124 = .64

   SE = sqrt((.5 * (1 - .5))/124) = 0.045

   z = (.64 - .50)/.045 = 3.12

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

3-4. p_val <- pnorm(3.12, 0, 1, lower.tail = FALSE) # p-value = 0.0009

5. Decision?

Parametric inference on means

Inference on means

1. From the central limit theorem, the distribution of sample means $\bar{x}$, has what shape?

- A: Normal!

2. And what value (symbol) is the sampling distribution of $\bar{x}$ center at?

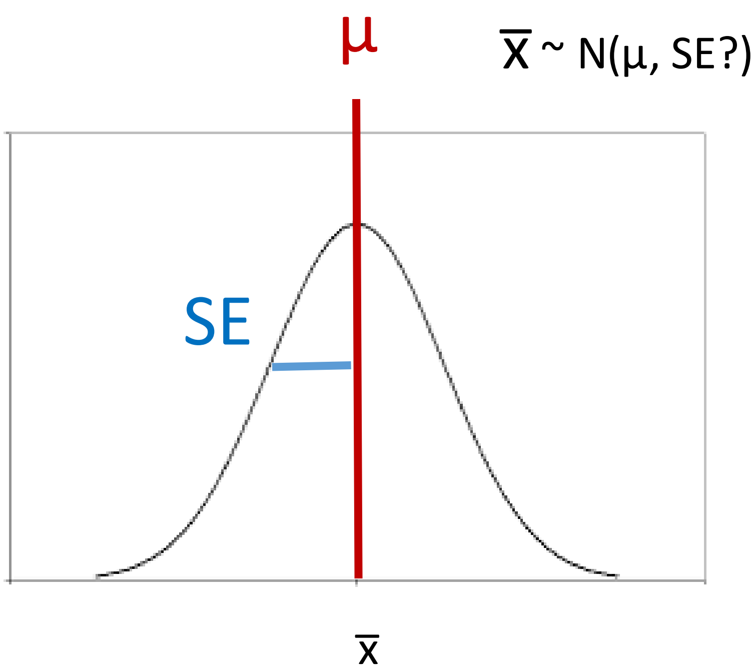
- A: μ

3. What other piece of information would be need to plot the sampling distribution of $\bar{x}$?

- A: SE

4. How can we estimate the SE if we had data?

- A: We could use the bootstrap, or…



Standard Error of Sample Means

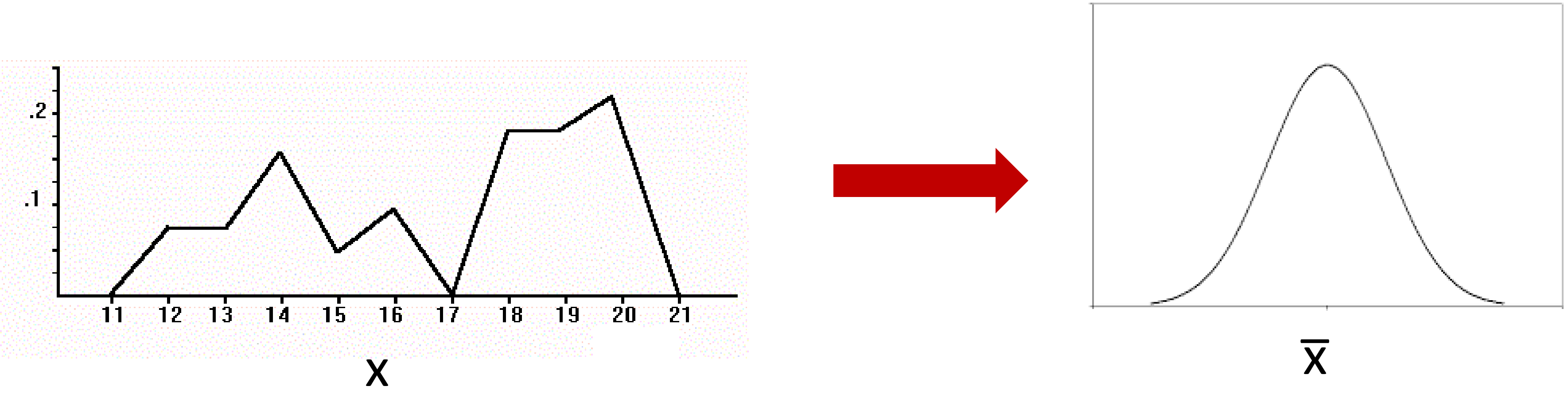
When choosing random samples of size n from a population with mean μ and standard deviation σ, the standard error of the sample means is:

$$SE = \frac{\sigma}{\sqrt{n}}$$

The larger the sample size (n), the smaller the standard error

Central Limit Theorem for Sample means

The sampling distribution of sample means ($\bar{x}$) *from **any population distribution*** will be normal, provided that the sample size is large enough



The more skewed the distribution, the larger sample size we will need for the normal approximate to be good

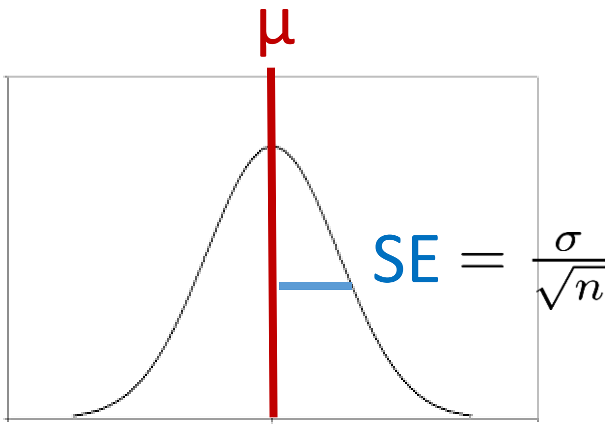
Sample sizes of 30 are usually sufficient. If the original population is normal, we can get away with smaller sample sizes

Central Limit Theorem for Sample means

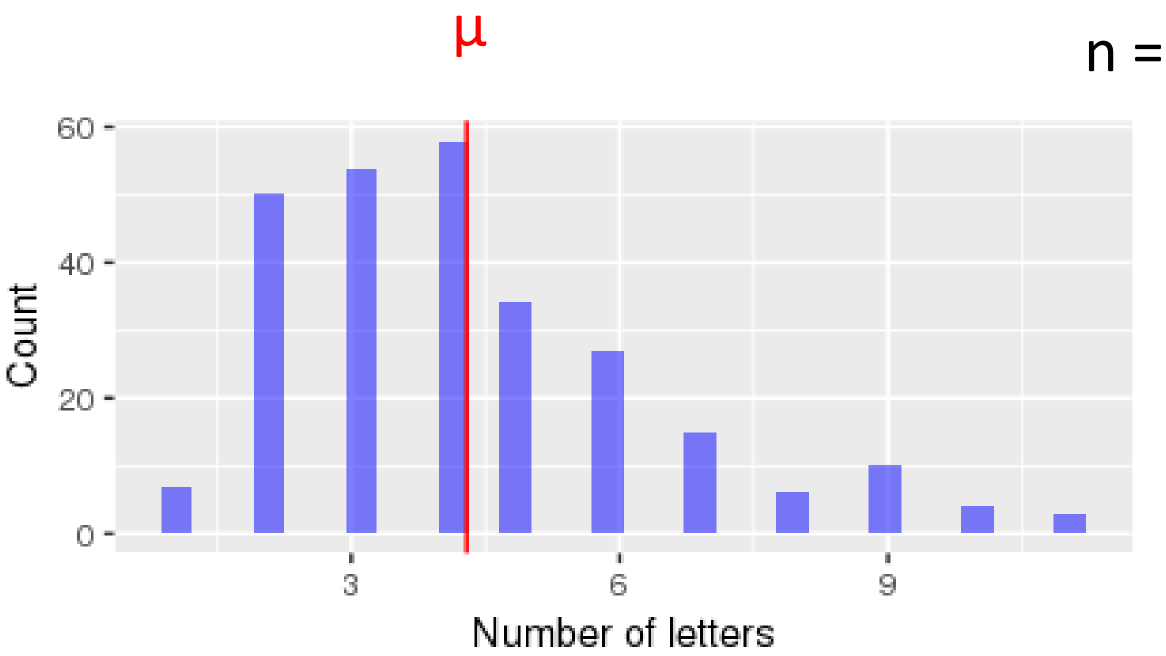
For random samples of size **n** from a population with mean **μ** and standard deviation **σ**…

the distribution of the sample means (**x̄**) is reasonably normal if the sample size is sufficiently large (n ≥ 30), with the mean **μ** and standard error $SE = \frac{\sigma}{\sqrt{n}}$

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



Gettysburg address word length sampling distribution



n = 10

10, 3, 3, 3, 4,
3, 2, 6, 10, 5

$\bar{x} = 5$

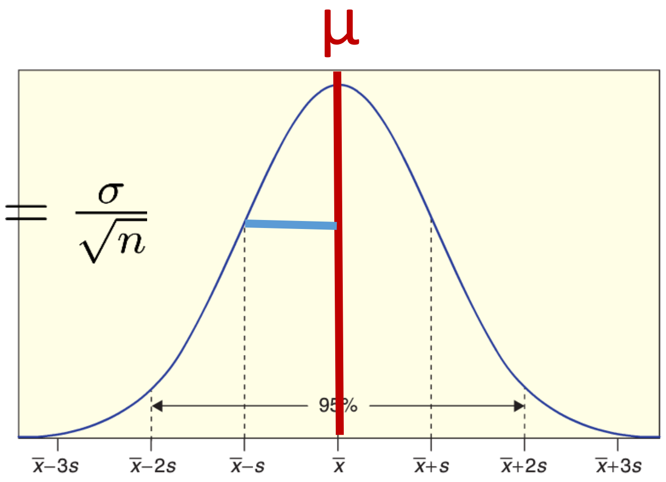
2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$



$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Sampling distribution!

Gettysburg sampling distribution app

Inference of the population mean μ

Now that we have a formula for the standard error $SE = \frac{\sigma}{\sqrt{n}}$
can we use it for inference on the population parameter μ?

1. Can we create confidence intervals for μ using: $\bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$

2. Can we run hypothesis tests to test H_0: μ = μ_0 using: $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

Unfortunately, we can’t because we don’t know σ

😢

Any ideas what we could do…?

Inference of the population mean μ

For proportions, ***we used our sample estimate of p̂ for the population parameter π*** and to compute the standard error, and the sampling distribution was still a normal distribution, so everything worked

$$\hat{p} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}) \qquad\qquad \hat{p} \ \pm \ z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Unfortunately, if we replace s for σ in our formulas, the sampling distribution of $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ is not a normal distribution

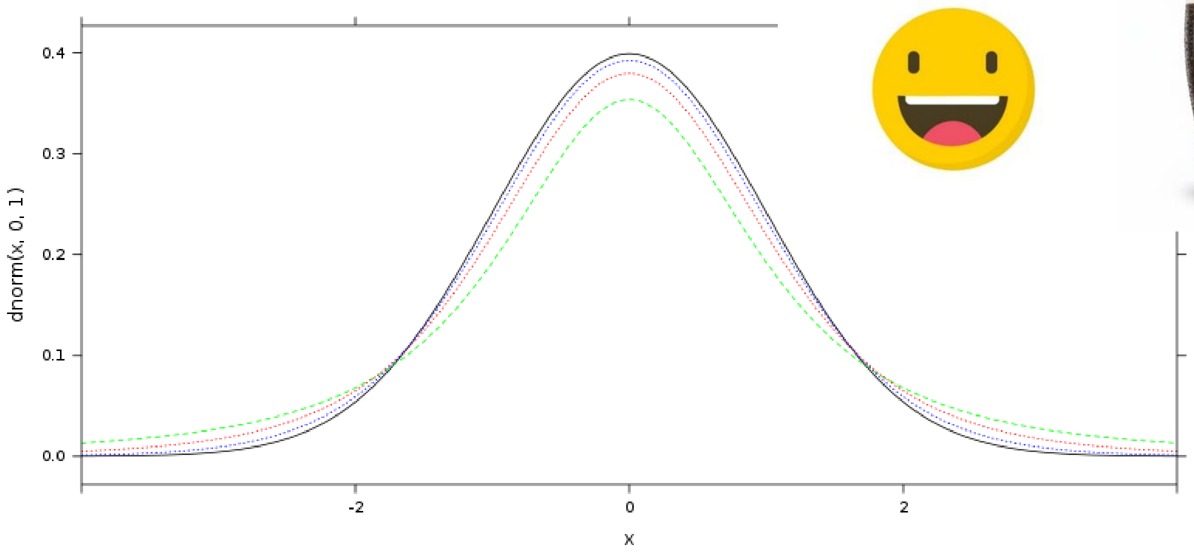
😢

Inference of the population mean μ

Fortunately, about 100 years ago, William Sealy Gosset figured out that this sampling distribution of $T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ is another parametric distribution called a t-distribution

The t-distribution is a parametric distribution that has one parameter called ***the degrees of freedom***

The t-distribution becomes more normal as the sample size *n* grows larger

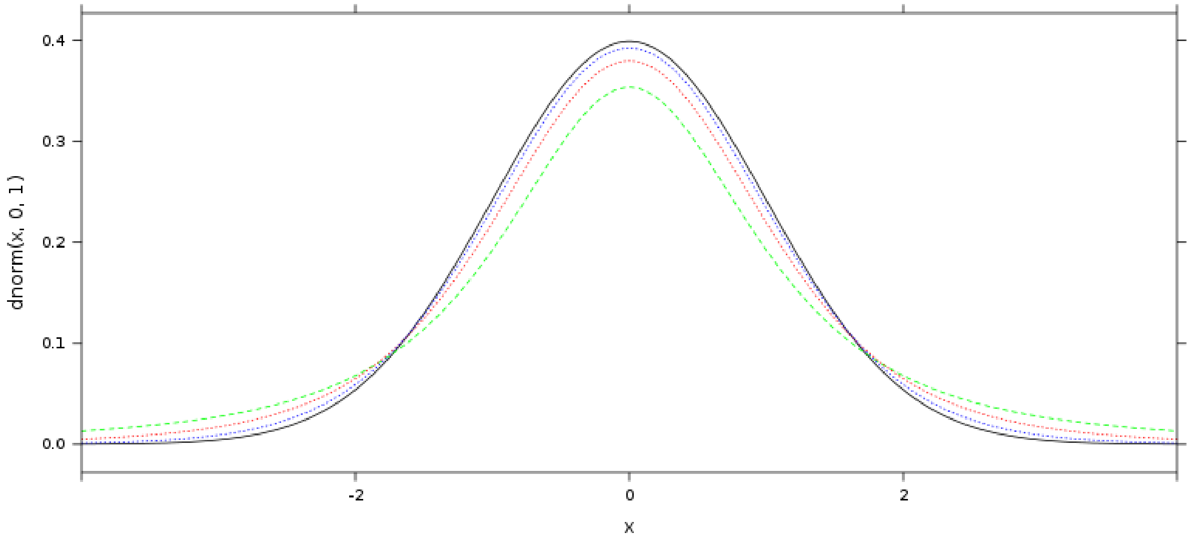


N(0, 1), df = 2, df = 5, df = 15

Inference using a t-distribution

When we have a sample of size n, the value of the degrees of freedom parameter is equal to n - 1

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$



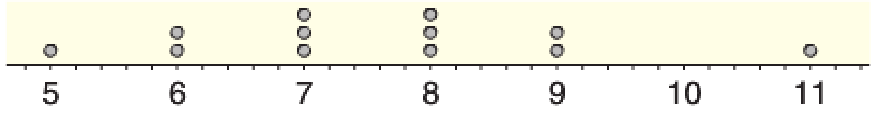
The fine print - this works if:

The underlying population has a distribution that is approximately normal or n > 30)

Is the t-distribution appropriate?

A

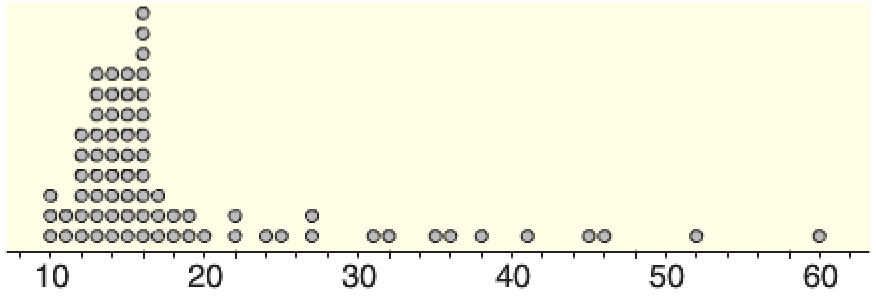
A sample with $n = 12$, $\overline{x} = 7.6$, and $s = 1.6$



Distribution seems normal
so OK to use t-distribution

B

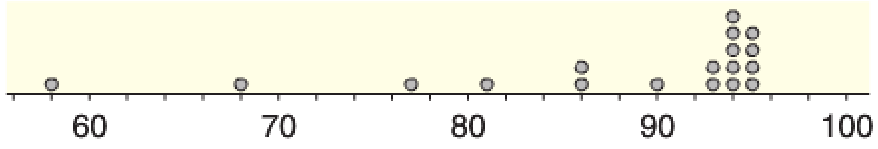
A sample with $n = 75$, $\overline{x} = 18.92$, and $s = 10.1$



Sample size is larger than n = 30
so OK to use the t-distribution

C

A sample with $n = 18$, $\overline{x} = 87.9$, and $s = 10.6$



Population distribution does not
look normal and n < 30 so NOT ok
to use the t-distribution

Calculating probabilities and quantiles from a t-distribution

We can use R to calculate probabilities and quantiles from the t-distribution

P(T ≤ t; deg_of_free)

Quantiles/critical values

Example: Suppose we have a sample size is n = 16, use R code to calculate the following values from a t-distribution:

1. The 2.5th and 97.5th percentiles
2. The probability that a t-statistic is more than 1.5
3. Then calculate these same values for the standard normal distribution

Let’s try it in R!

Calculating probabilities and quantiles from a t-distribution

If a sample size is n = 16, calculate:

1. Calculate the 2.5th and the 97.5th percentiles
2. Find the probability that a t-statistic is more than 1.5
3. Calculate these same values for the standard normal

1. `qt(c(.025, .975), df=15)` = [-2.13 2.13]

2. `pt(1.5, df=15, lower.tail = FALSE)` = 0.077

3. A. `qnorm(c(.025, .975), 0, 1)` = [-1.96 -1.96]

   B. `pnorm(1.5, 0, 1, lower.tail = FALSE)` = 0.067

Parametric confidence intervals for a single mean

Confidence Interval for a single mean

For a normally distributed variable (e.g., a proportion), we saw that we could create a confidence interval with the formula:

Sample statistics ± z* × SE

We can use a similar formula for creating a confidence interval for μ using a t-distribution and our standard error formula: $SE = \frac{s}{\sqrt{n}}$

A confidence interval for μ is: x̄ ± t* × SE $= \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$

Where t* is an endpoint chosen from a t-distribution with n-1 df to give the desired confidence level

The t-distribution is appropriate if the distribution of the population is approximately normal or the sample size is large (n >= 30)

How many birds to cats kill?

A study by Loyd et al (2013) in Biological Conservation, used KittyCams to record all activity of n = 55 domestic cats that hunt outdoors

The video footage showed that the mean number of kills per week for these cats was 2.4 with a standard deviation of 1.51

Find and interpret a 99% confidence interval for the mean number of kills per week by US household cats that hunt outdoors

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Let’s try it in R!

How many birds to cats kill?

What are the values of the following?

- $\bar{x}$ = 2.4
- n = 55
- s = 1.51
- t^*_{99} = qt(.995, 54) = 2.67

Plugging in to our formula we get:

- 2.4 ± qt(.995, 54) * 1.51/sqrt(55)
- [1.86 2.94]

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Hypothesis tests for a single mean

Parametric test for a single mean μ

When the distribution of a statistic under H_0 is **normal**, we computed a standardized test statistic using:

$$z = \frac{Sample\ \ Statistic\ \ -\ \ Null\ \ Parameter}{SE}$$

When testing hypotheses for a single mean we have:

- H_0: $\mu = \mu_0$ (where μ_0 is specific value of the mean)

Thus, the null parameter is μ_0, and the sample statistics is $\bar{x}$ …

If we could use $SE = \dfrac{\sigma}{\sqrt{n}}$

$$z = \frac{\bar{x} - \mu_0}{SE}$$

Parametric hypothesis tests for a single mean μ

We can estimate the standard error using $SE = \frac{s}{\sqrt{n}}$

However, this makes the standardized statistic follow a t-distribution with n-1 degrees of freedom rather than a normal distribution

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

This works if n is large or the data is reasonably normally distributed

Because we are using a t-distribution to find the p-value, this is called a **t-test**

t-test for a single mean

To test:

H_0: $\mu = \mu_0$ vs.

H_A: $\mu \neq \mu_0$ (or a one-tailed alternative)

We use the t-statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

A p-value can be computed using a t-distribution with n-1 degrees of freedom

- Provided that the population is reasonable normal (or the sample size is large)

The Chips Ahoy! Challenge

In the mid-1990s a Nabisco marking campaign claimed that there were at least 1000 chips in every bag of Chips Ahoy! cookies

A group of Air Force cadets tested this claim by dissolving the cookies from 42 bags in water and counting the number of chips

They found the average number of chips per bag was 1261.6, with a standard deviation of 117.6 chips

Test whether the average (mean) number of chips per bag is greater than 1000. Do the results confirm Nabisco's claim?

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

pt(t, df = deg_of_free)

Let's try it in R!

The Chips Ahoy! Challenge

H_0: μ = 1000 vs H_A: μ > 1000

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$\overline{x}$ = 1261.6

s = 117.6

n = 42

df = 41

SE = 117.6/sqrt(42)

t = (1261.6 – 1000)/18.141 = 14.42

P-value: pt(14.32, df = 41)   < 10^-16

Does this verify chips ahoy!'s claim?

John Tukey quote

“Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question”

- *The future of data analysis*. Annals of Mathematical Statistics 33 (1), (1962), page 13.