INFSCI 2595

Fall 2019

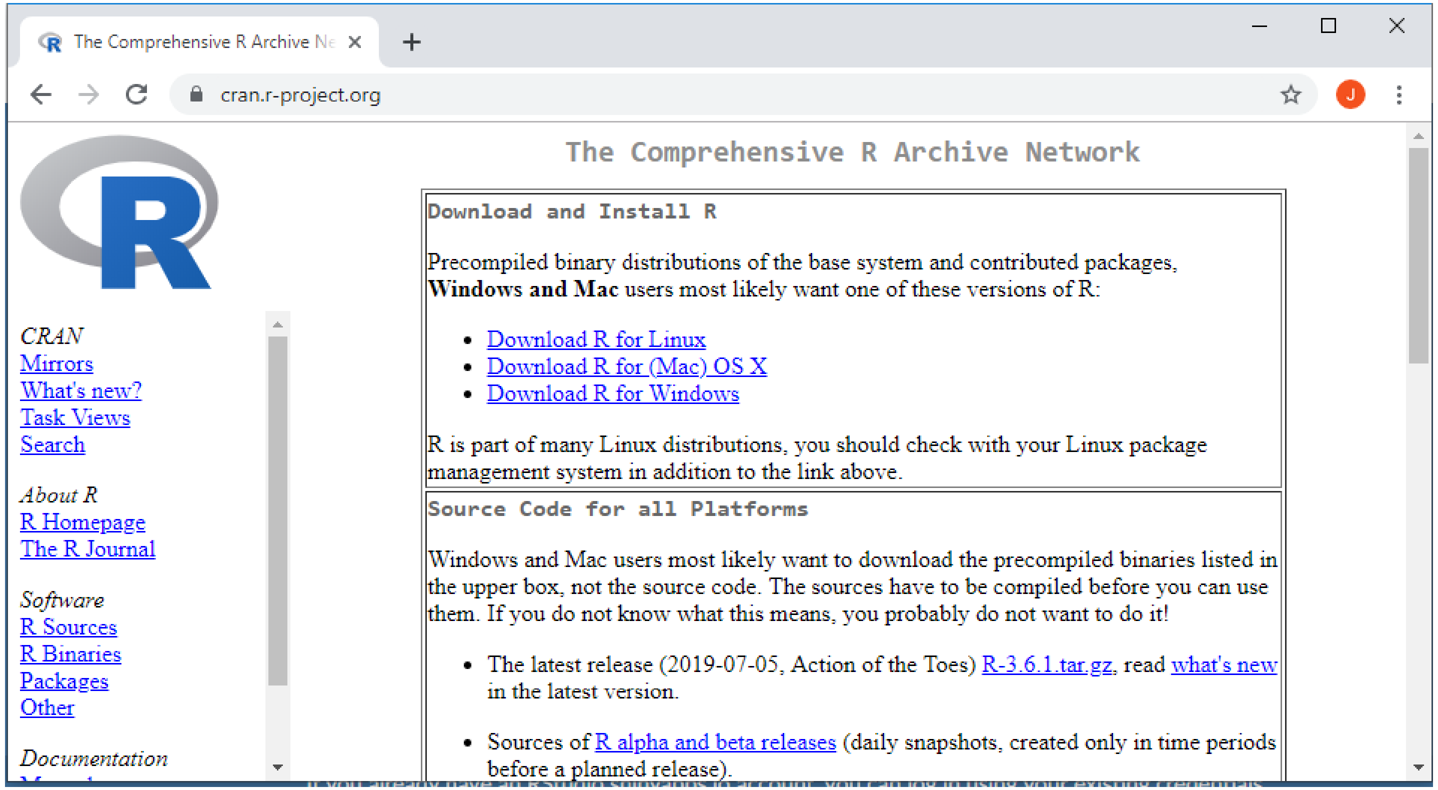
ROOM LOCATION

Lecture 02

Introduction to the R ecosystem

Download and install R

- Go to the CRAN:
- https://cran.r-project.org/

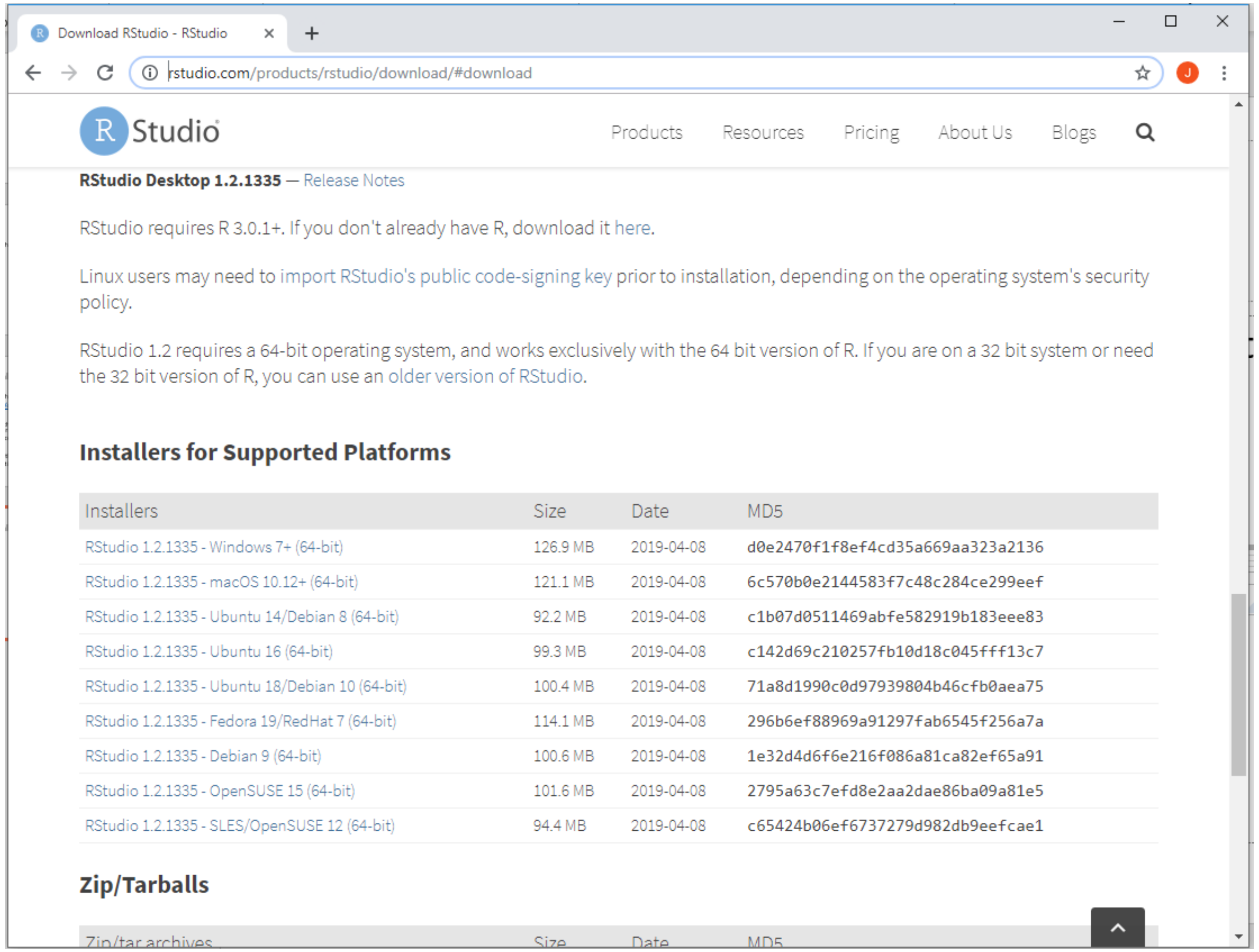


- Select the Download link for your operating system of choice.

- Follow the onscreen instructions.

Download and install RStudio

- Go to the following link:
- https://www.rstudio.com/products/rstudio/download/#download
- Select the installer for your operating system of choice.
- Follow the onscreen instructions



Cool stuff in R – interactive web applications

- R Shiny allows creating fully interactive visualizations, models, and reports.
  - Essentially converts R to HTML!

- See https://shiny.rstudio.com/ for examples and a gallery of user created apps.

Quick in-class example

Probability review

We will introduce probability in terms of…

Pitt vs Penn State football!!

Source: https://heinzfield.com/football/pittsburgh-panthers/

Source: https://www.deviantart.com/donatello16/art/We-are-Penn-State-763249687

Problem statement

- Assumptions:
  - Pitt and Penn State (PSU) are the only college football teams that fans care about in Pennsylvania (PA).
  - We can use attendance to the Pitt vs PSU games to represent the population of PA.
  - We can treat each fan at the game as being independent of all other fans (not true in reality).

- **Objective: Determine the probability that a PA college football fan is a Pitt fan.**

We will assume that we can represent this situation with 2 random variables

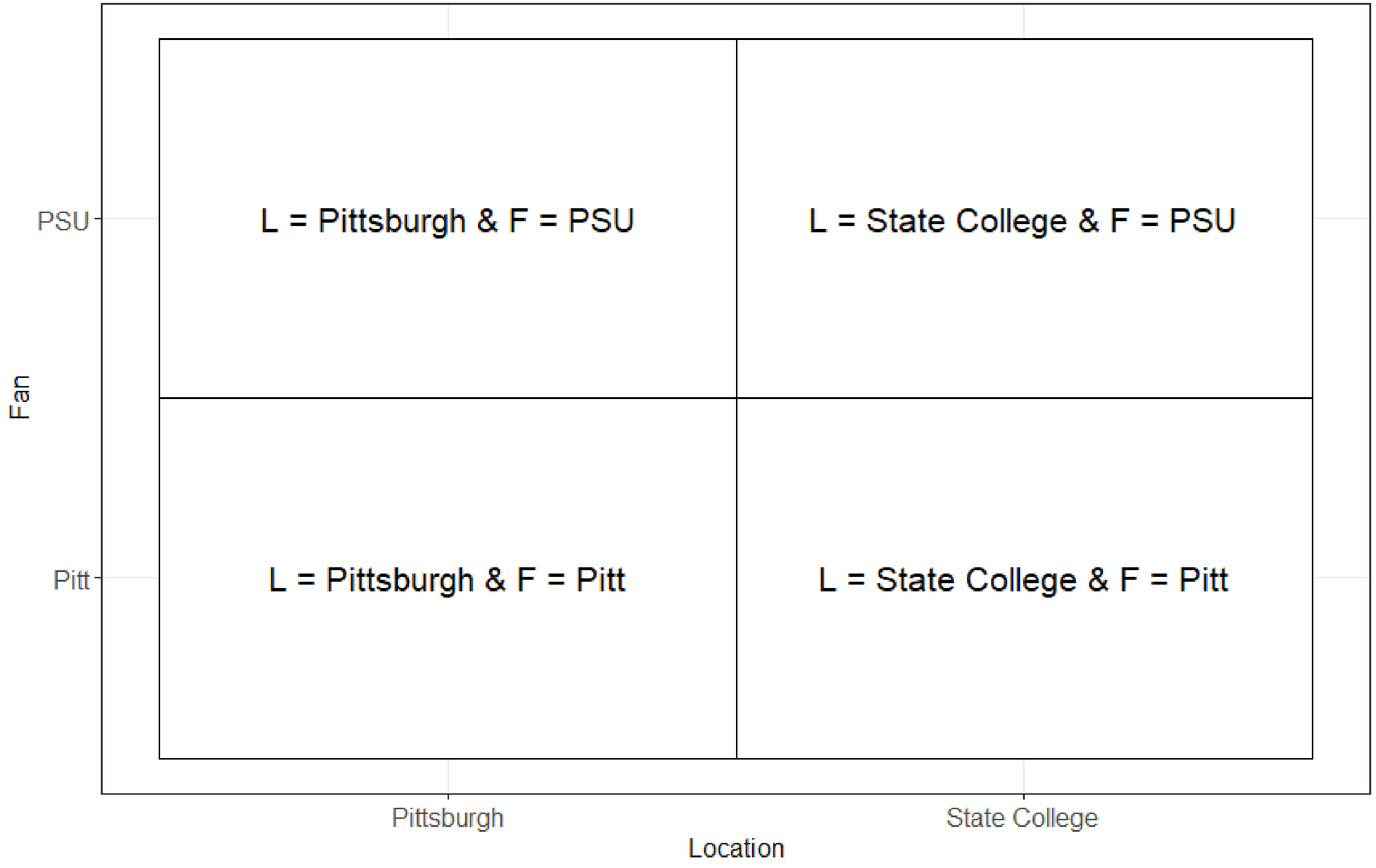
- `Location` – denote as `L`
  - The game can be played in either `Pittsburgh` (Pitt's home field) or `State College` (PSU's home field).

- `Fan` – denote as `F`
  - A college football fan in PA can be a fan of either `Pitt` or `PSU`.

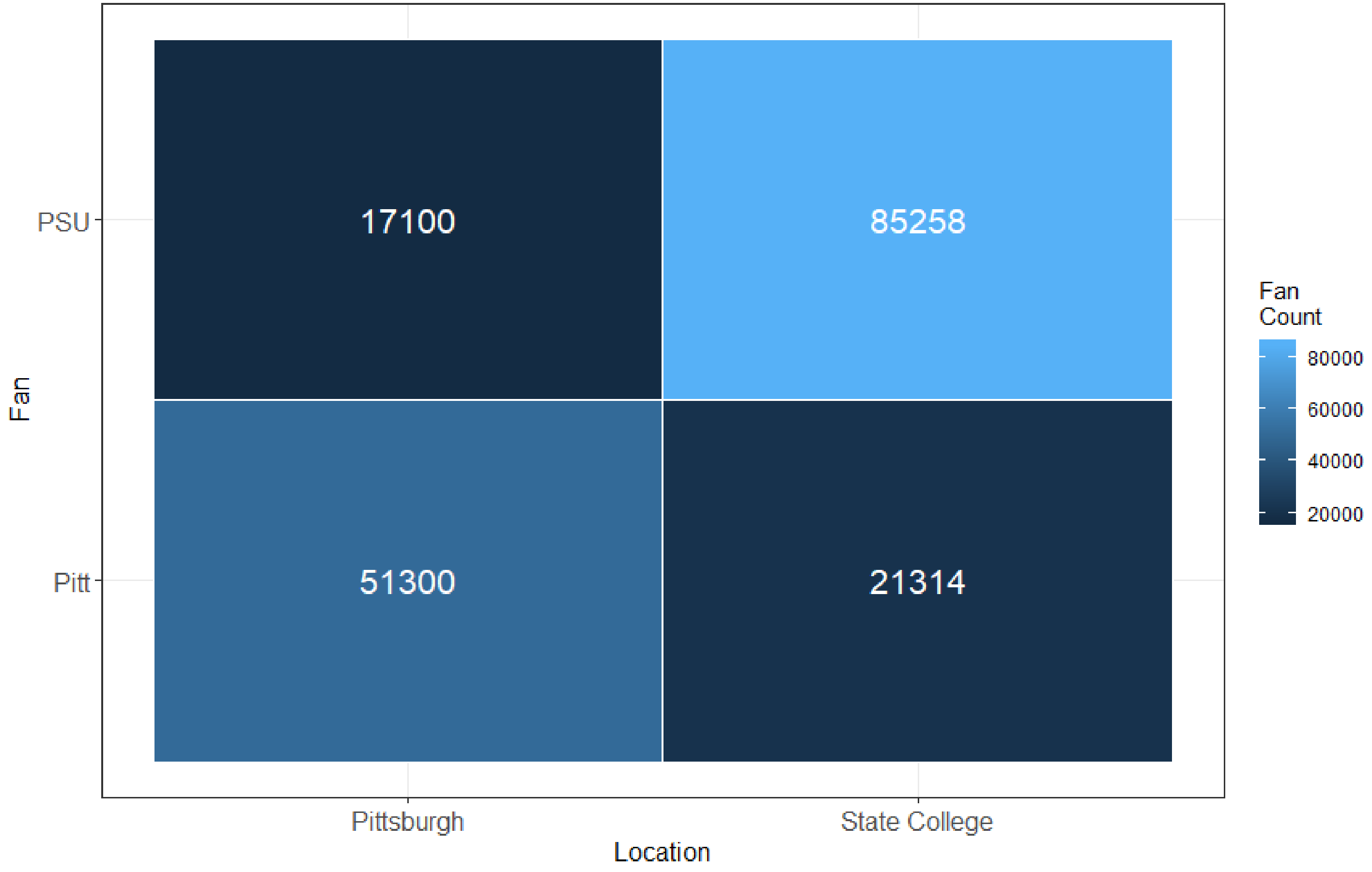
4 possible combinations of these 2 random variables

- `L` = `Pittsburgh` & `F` = `Pitt`
- `L` = `Pittsburgh` & `F` = `PSU`
- `L` = `State College` & `F` = `Pitt`
- `L` = `State College` & `F` = `PSU`

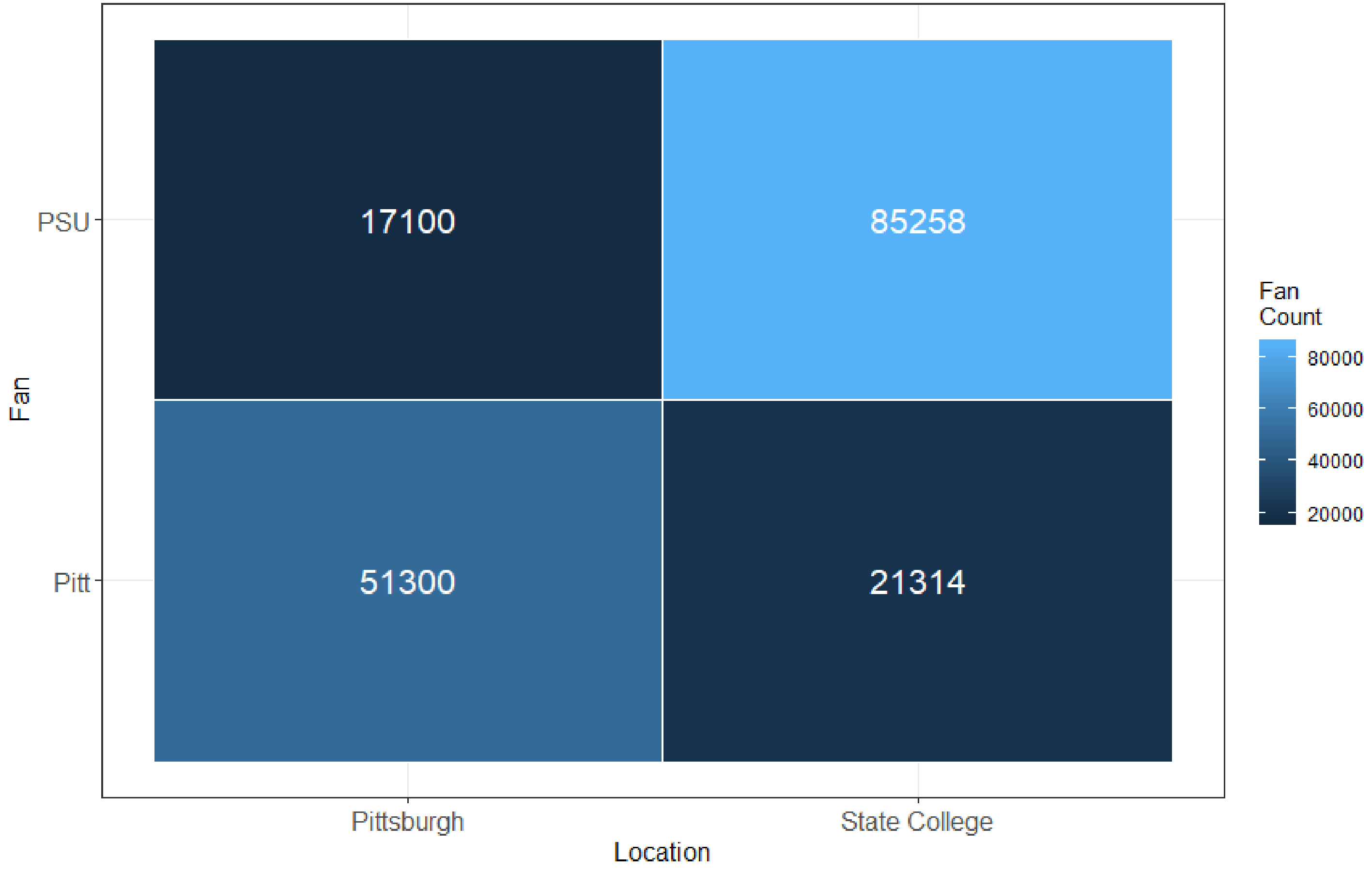
Visualize the 4 combinations as a matrix



Include some example values for the number of observations per combination



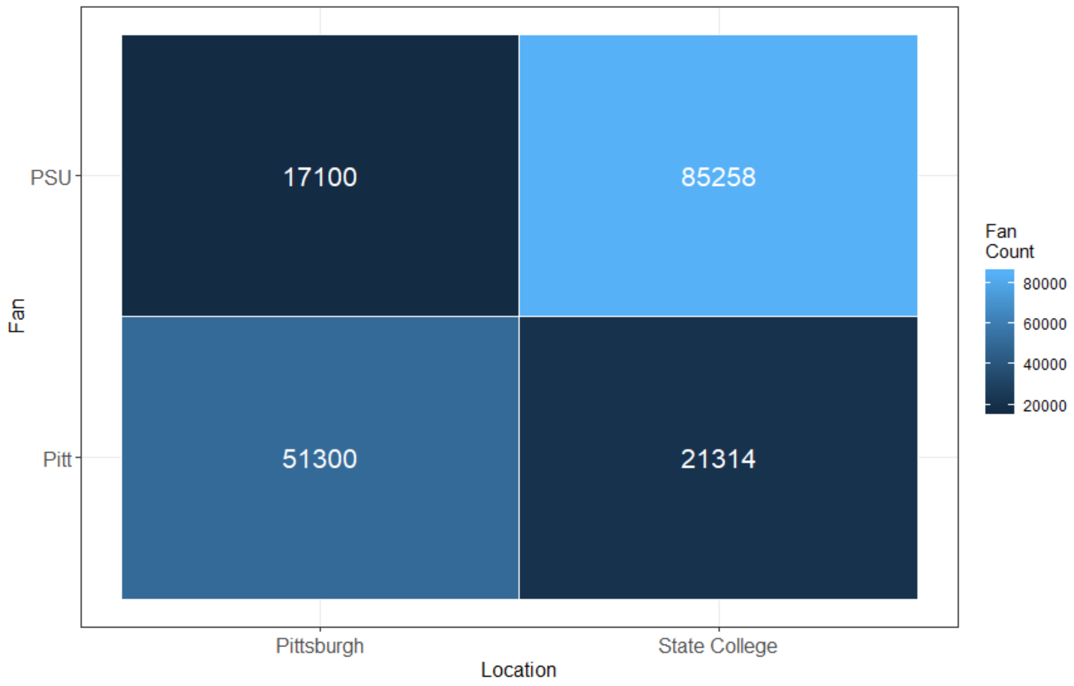
To meet our objective, we need to first derive several basic probability rules



Start with the basic definition of probability

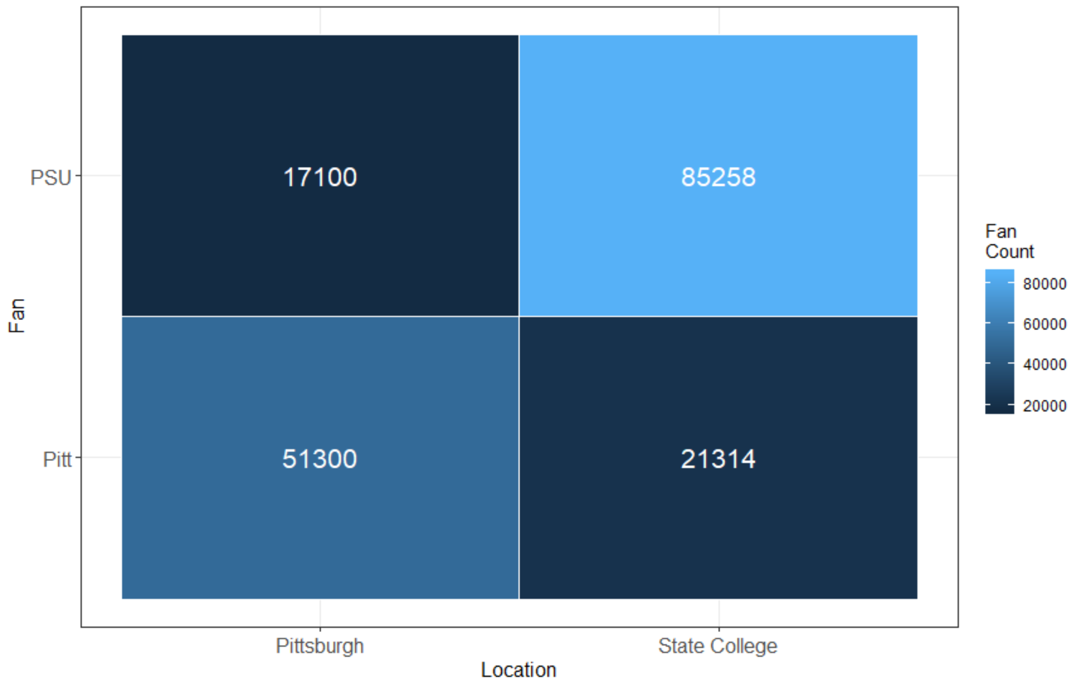
- The probability of an event is the fraction of times that event occurs out of the total number of trials, in the limit that the total number of trials goes to infinity.

In our example, an event is the combination of `Location` and `Fan`



- How do we calculate the probability of each event?

In our example, an event is the combination of `Location` and `Fan`

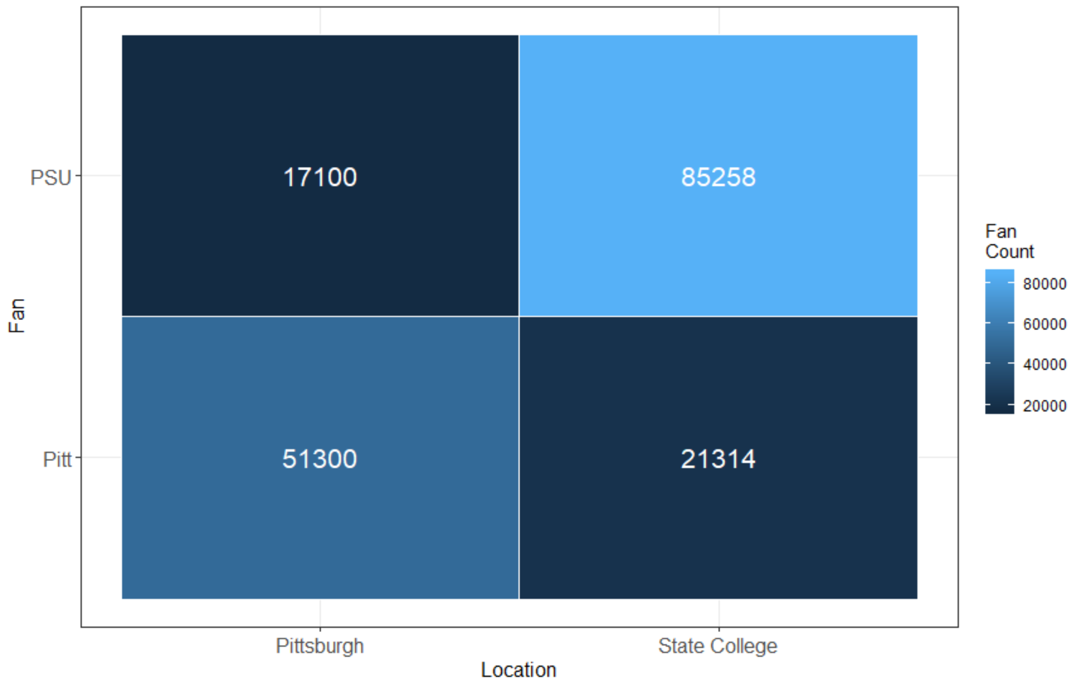


- How do we calculate the probability of each event?

$$P(\text{event}) = \frac{\text{number of times event occurs}}{\text{total number of trials}}$$

- Total number of trials in our example?

In our example, an event is the combination of `Location` and `Fan`



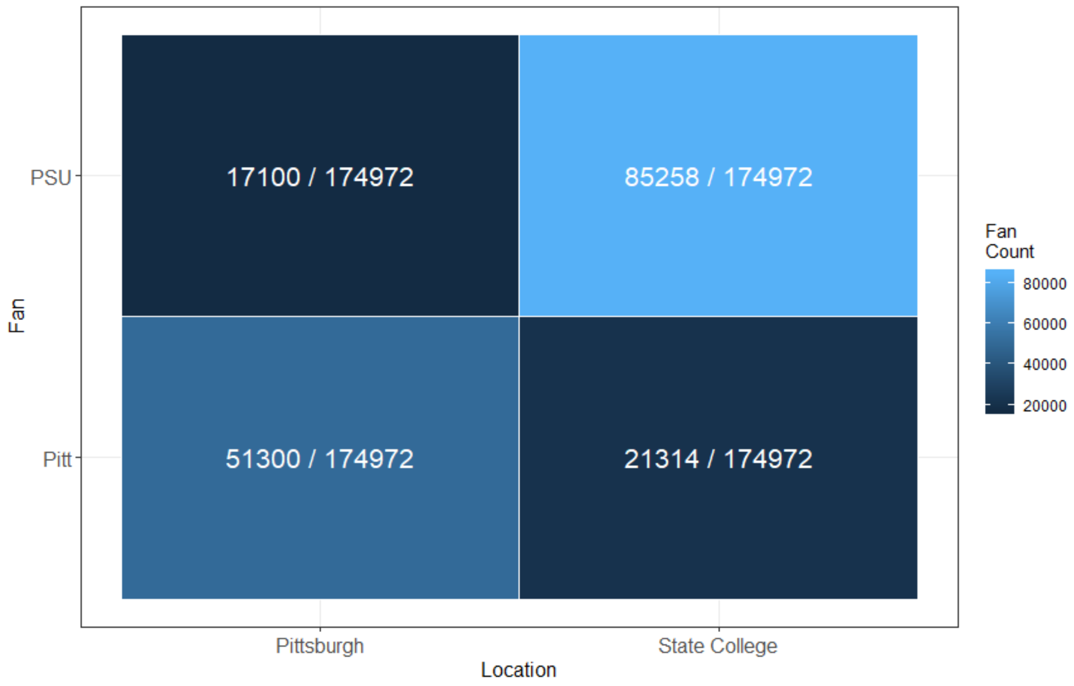
- How do we calculate the probability of each event?

$$P(\text{event}) = \frac{\text{number of times event occurs}}{\text{total number of trials}}$$

- Total number of trials in our example?

- Sum up the counts across the 4 combinations!

In our example, an event is the combination of `Location` and `Fan`



- How do we calculate the probability of each event?

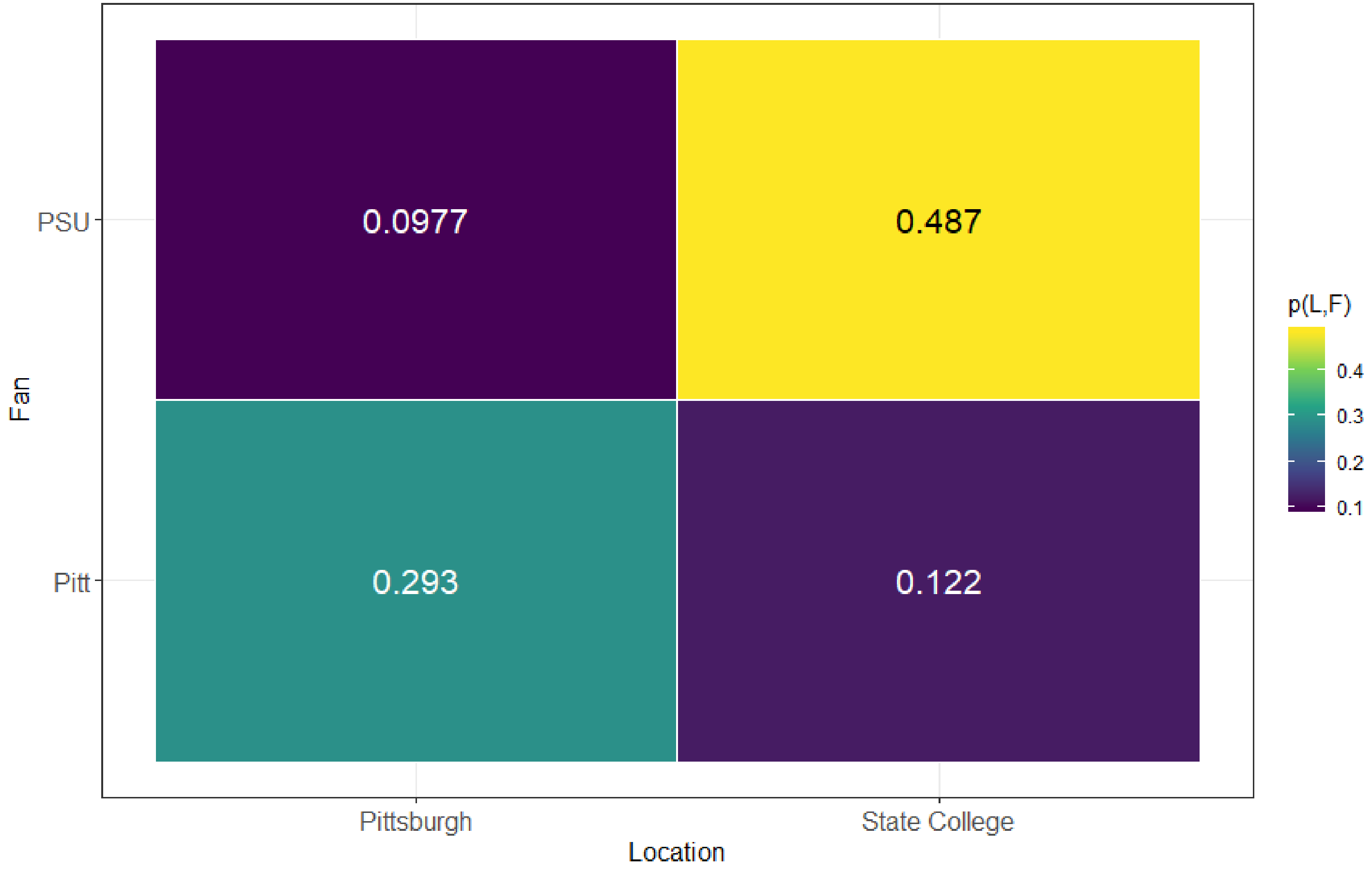
$$P(\text{event}) = \frac{\text{number of times event occurs}}{\text{total number of trials}}$$

- Total number of trials in our example?

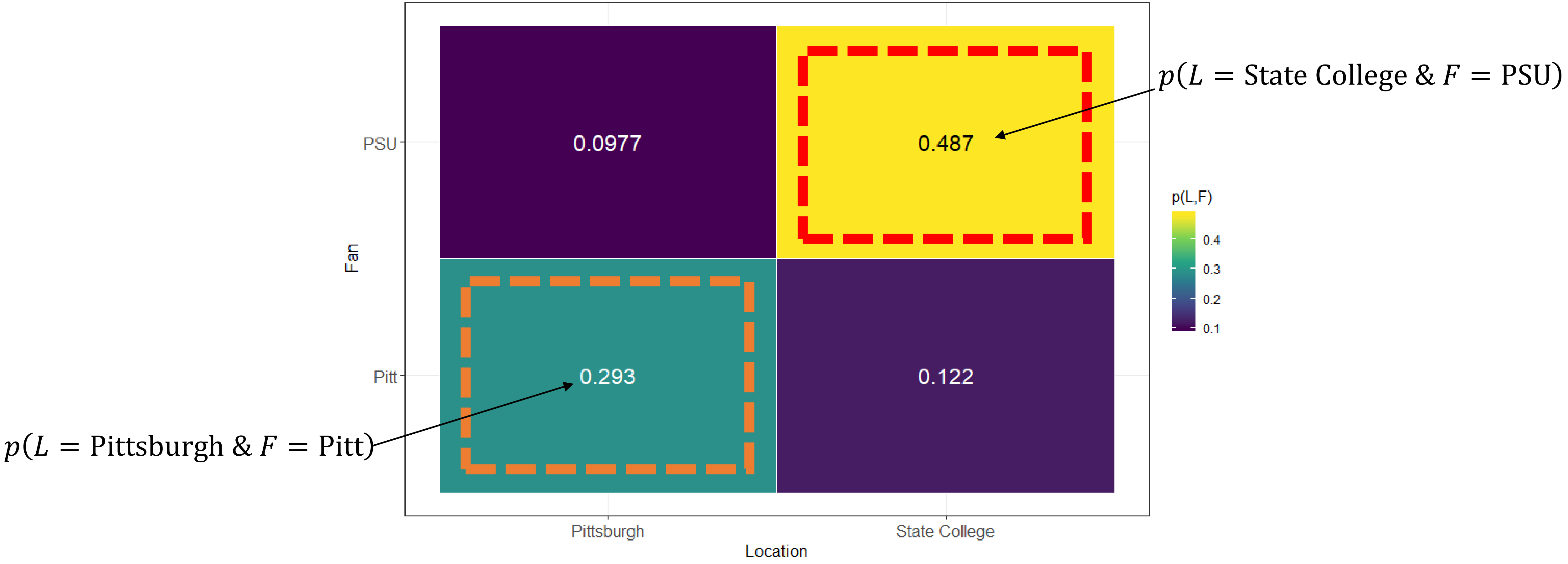
- Sum up the counts across the 4 combinations!
- Total number of trials = 174972

For now, allow 174972 to be represent approaching infinity…

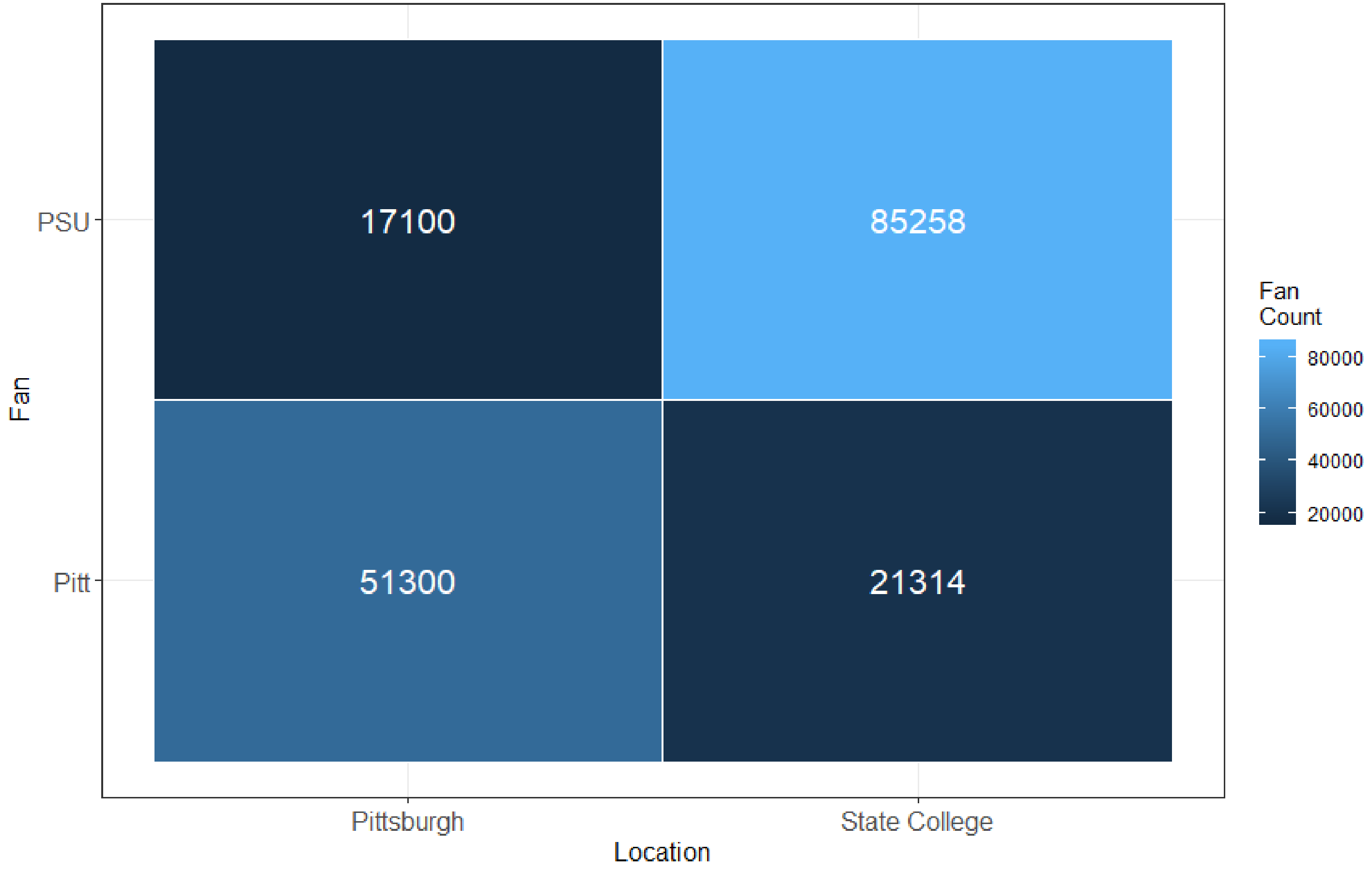
The probability of each **combination** is the referred to as the **joint probability**



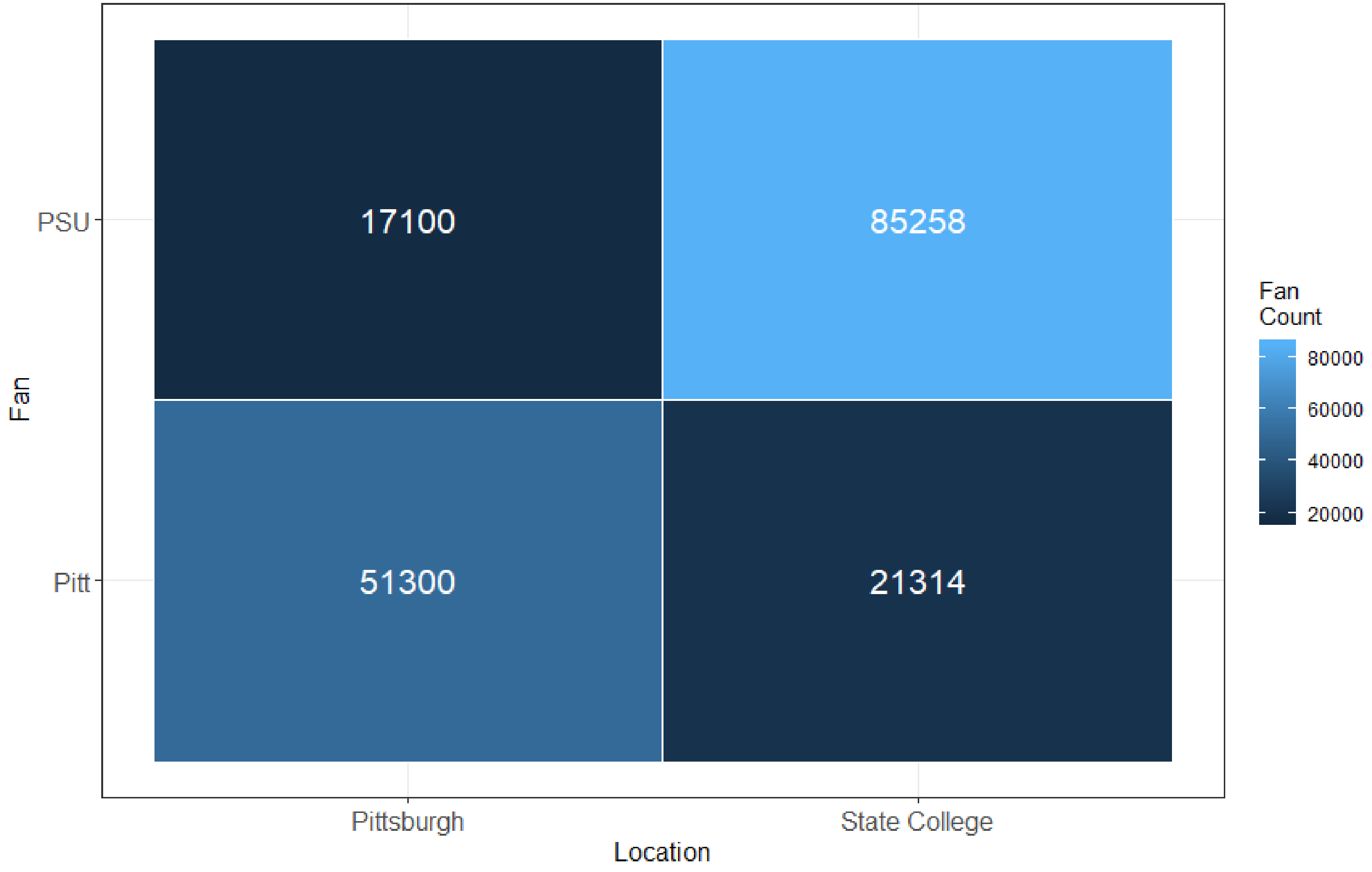
The probability of each **combination** is the referred to as the **joint probability**



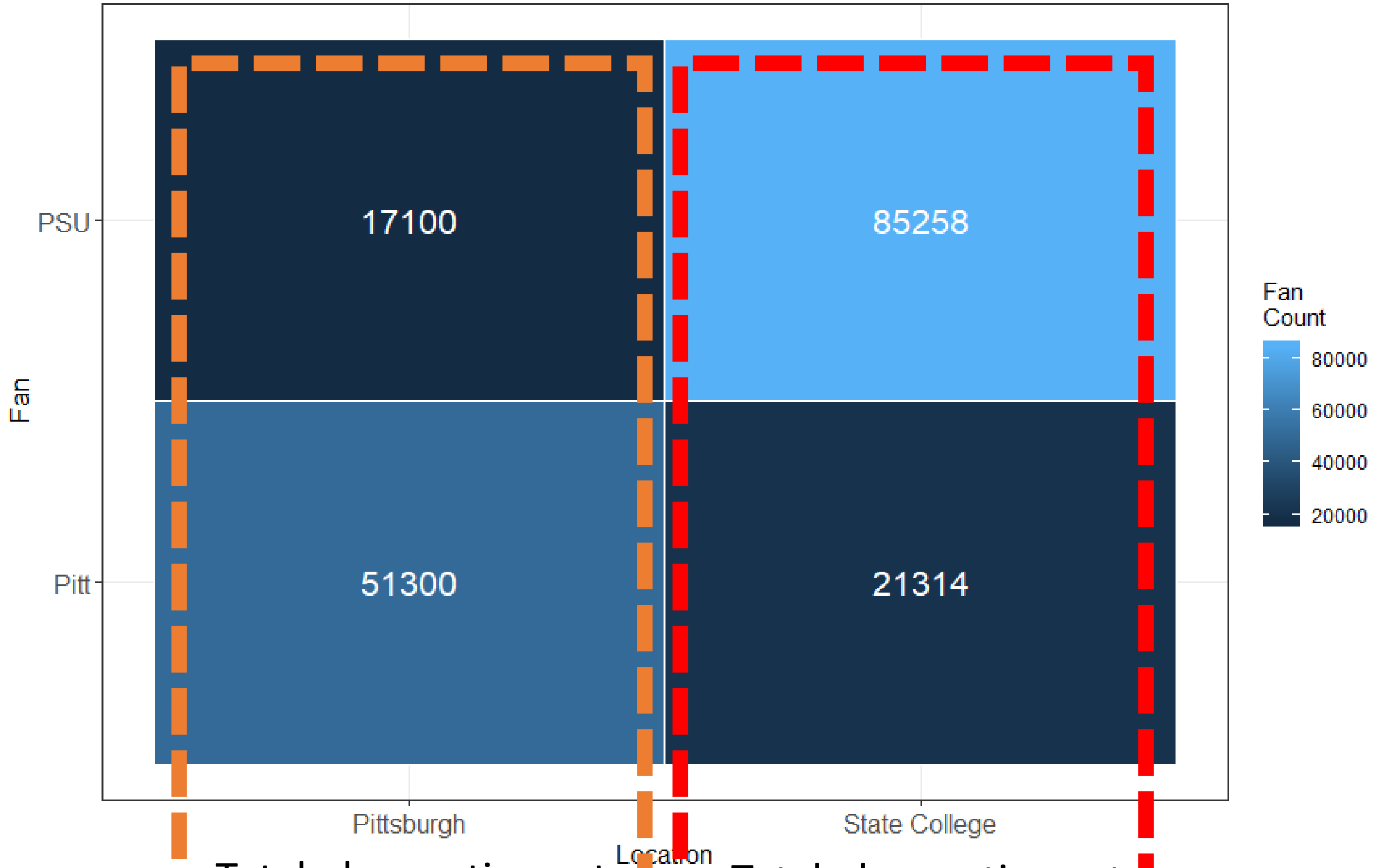
However, rather than working with the joint probabilities, we will work with the counts directly.



How would we calculate $p(L = \text{Pittsburgh})$, irrespective of the value of `Fan`?



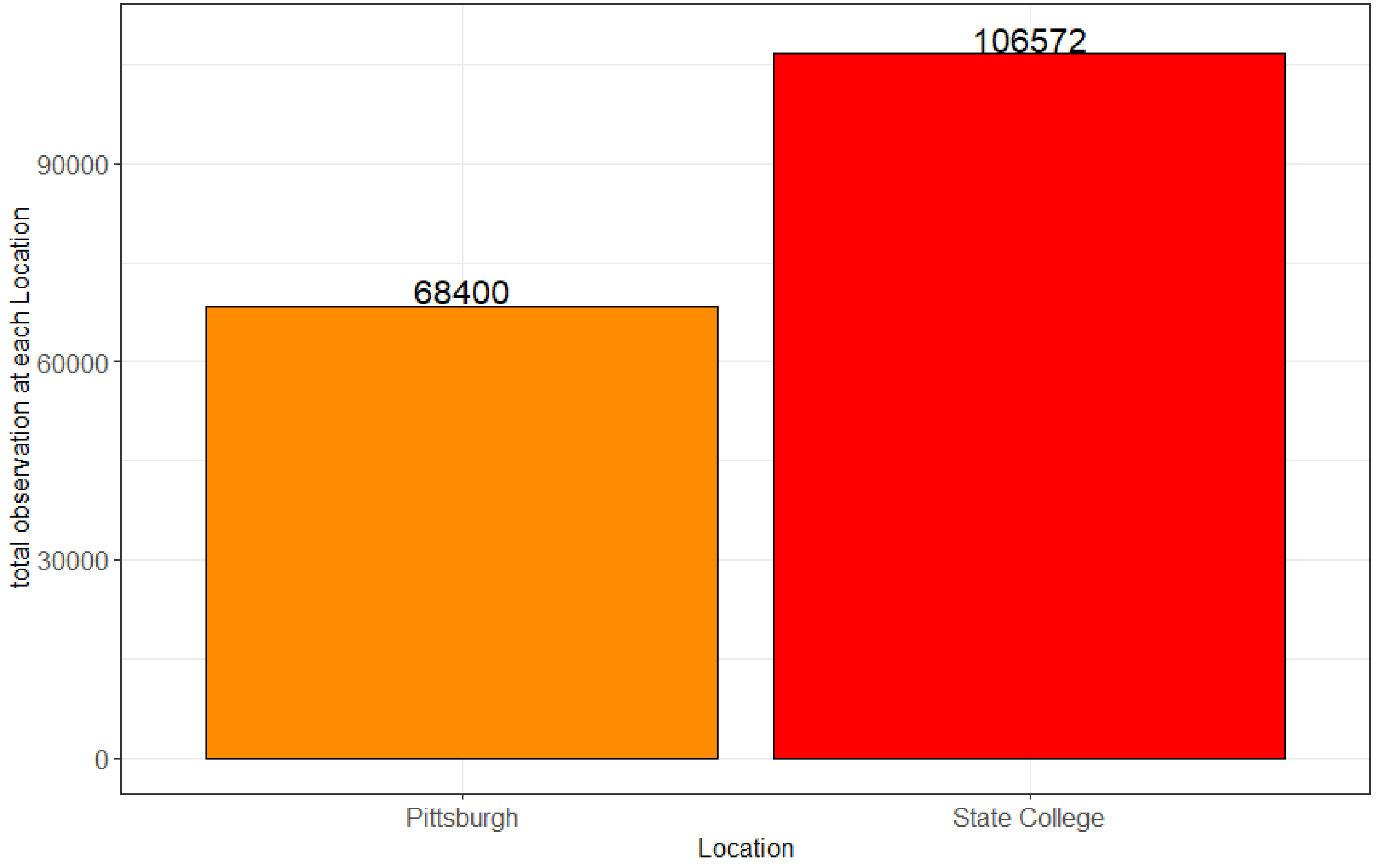
How would we calculate $p(L = \text{Pittsburgh})$, irrespective of the value of `Fan`?



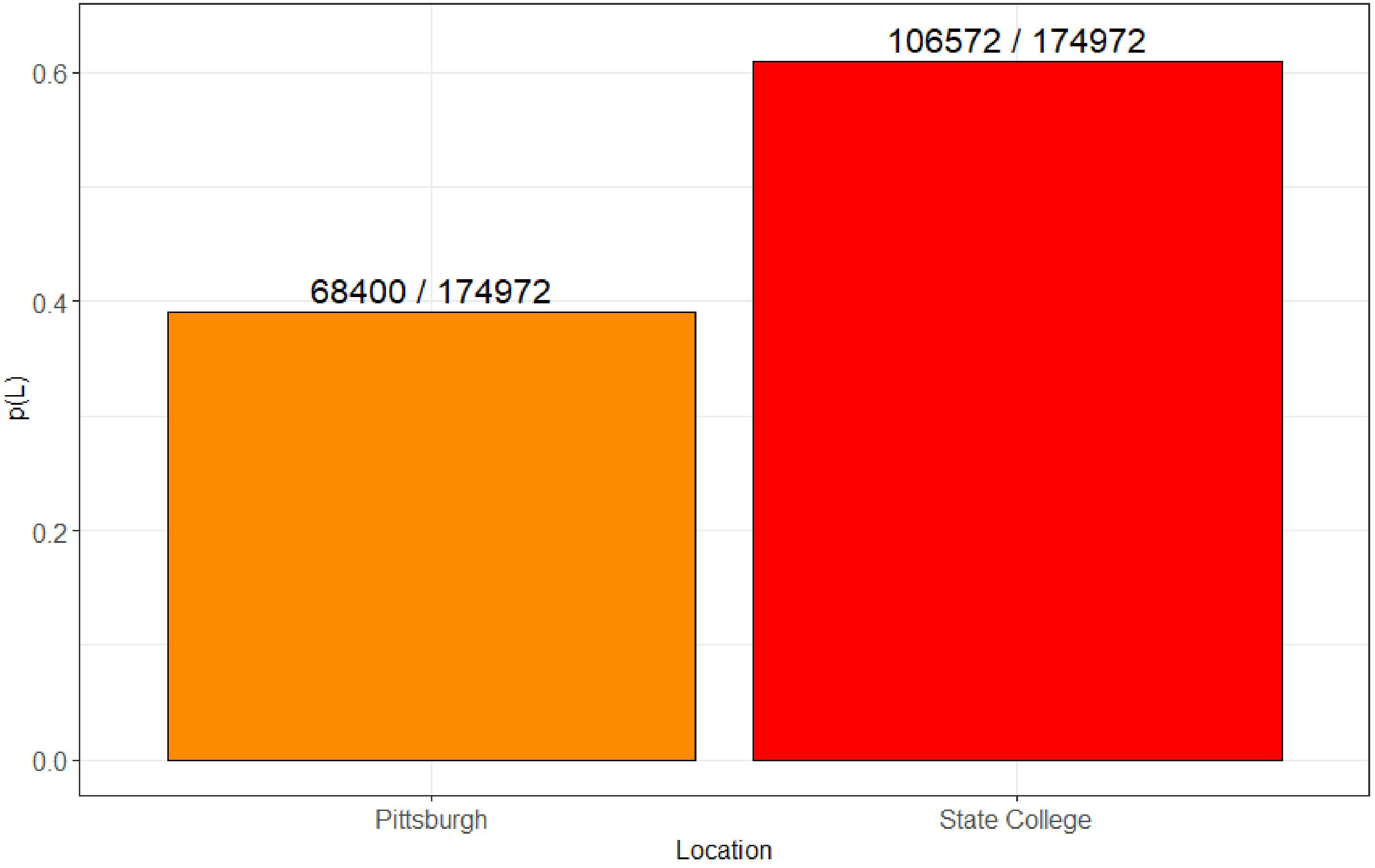
Total observations at $L = \text{Pittsburgh}$

Total observations at $L = \text{State College}$

Sum up the counts at each `Location`



Divide the total observations at each `Location` by the total number of trials



Work through the steps with the example values…

- Sum up the counts:

$$68400 = 51300 + 17100$$

- Divide by the total number of trials:

$$\frac{68400}{174972} = \frac{51300}{174972} + \frac{17100}{174972}$$

Work through the steps with the example values…

- Sum up the counts:

$$68400 = 51300 + 17100$$

- Divide by the total number of trials:

$$\frac{68400}{174972} = \boxed{\frac{51300}{174972}} + \boxed{\frac{17100}{174972}}$$

What are these quantities?

Work through the steps with the example values…

- Sum up the counts:

$$68400 = 51300 + 17100$$

- Divide by the total number of trials:

$$\frac{68400}{174972} = \boxed{\frac{51300}{174972}} + \boxed{\frac{17100}{174972}}$$

$p(L = \text{Pittsburgh} \;\&\; F = \text{Pitt})$

$p(L = \text{Pittsburgh} \;\&\; F = \text{PSU})$

Joint probabilities!

Work through the steps with the example values…

- Sum up the counts:

$$68400 = 51300 + 17100$$

- Divide by the total number of trials:

$$\boxed{\frac{68400}{174972}} = \boxed{\frac{51300}{174972}} + \boxed{\frac{17100}{174972}}$$

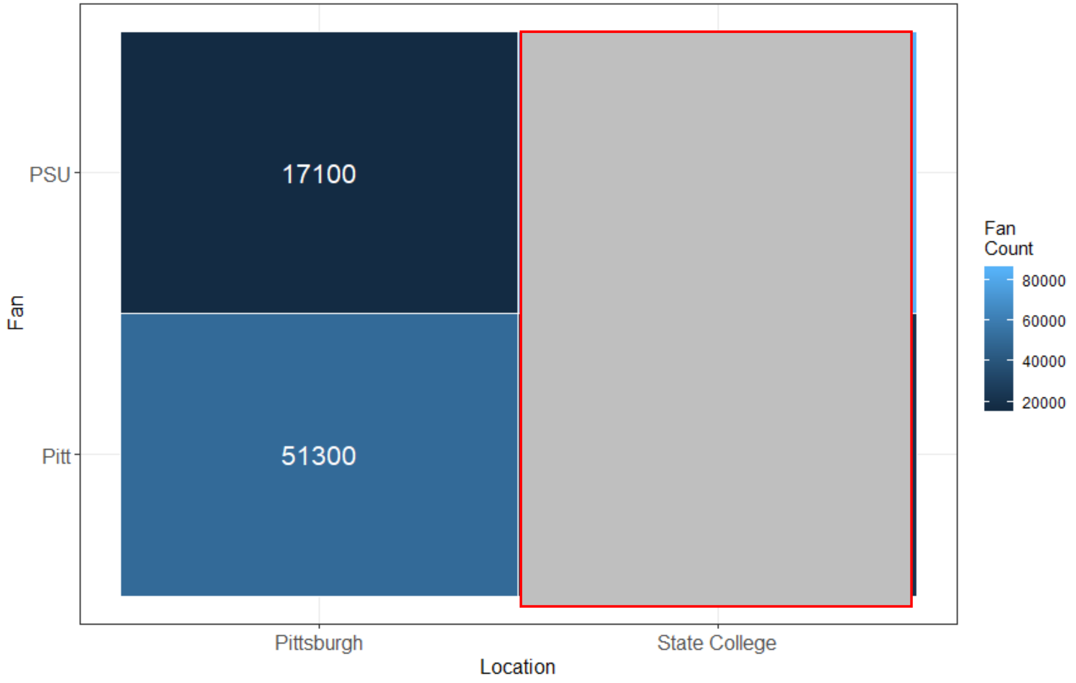
$$p(L = \text{Pittsburgh}) = p(L = \text{Pittsburgh}\ \&\ F = \text{Pitt}) + p(L = \text{Pittsburgh}\ \&\ F = \text{PSU})$$

The **sum rule** of probability -> *marginalize* or sum out the other variables

Next, focus on $L = \text{Pittsburgh}$. What is the probability of each `Fan` **GIVEN** the `Location`?

- The basic definition of probability remains the same.

- Has the total number of trials changed?

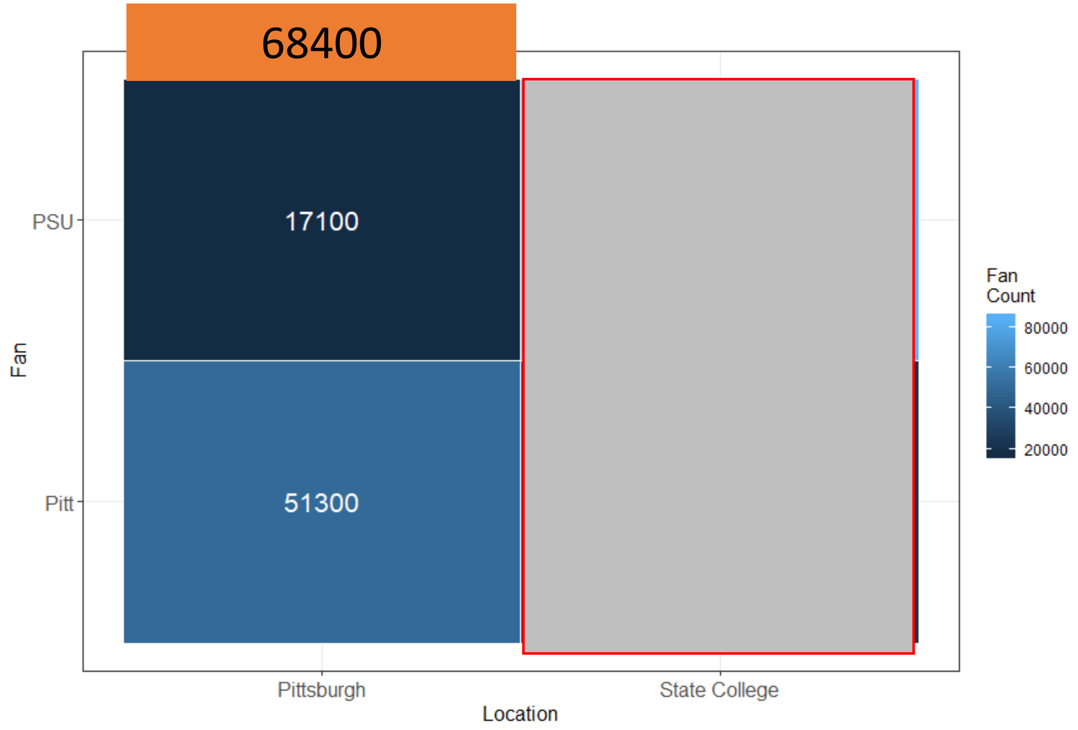


Next, focus on $L = \text{Pittsburgh}$. What is the probability of each `Fan` **<u>GIVEN</u>** the `Location`?

- The basic definition of probability remains the same.

- Has the total number of trials changed?
  - YES!

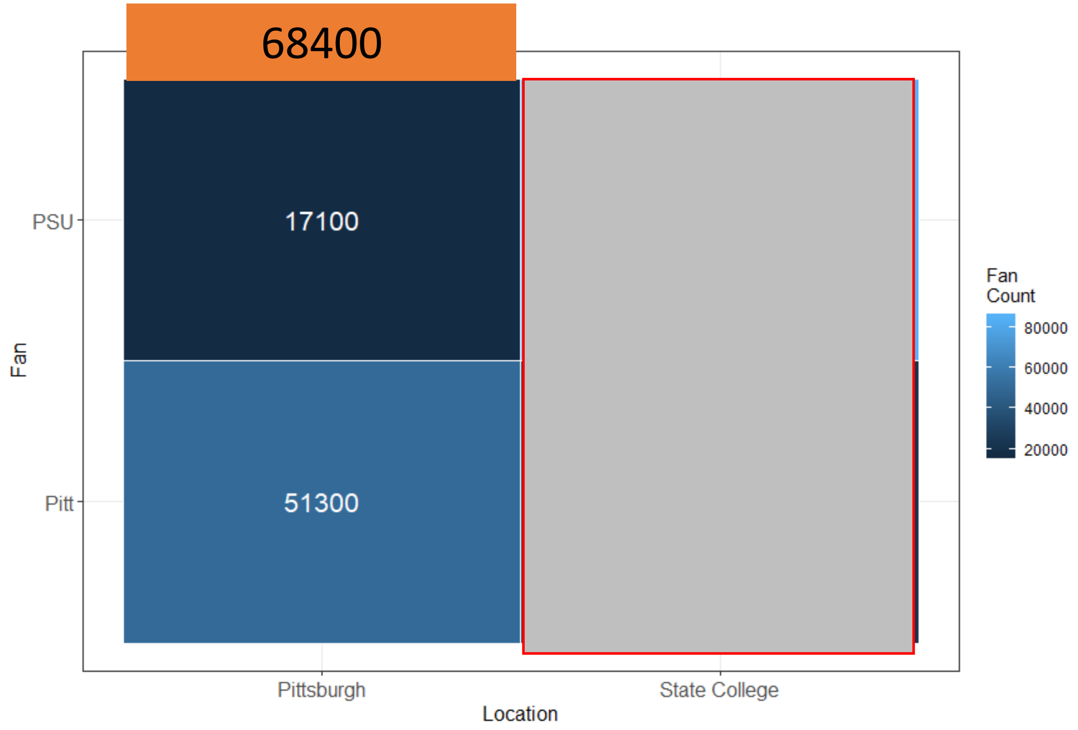
$$p(F = \text{Pitt}|L = \text{Pittsburgh}) = \frac{51300}{68400}$$



Next, focus on L = Pittsburgh. What is the probability of each `Fan` **<u>GIVEN</u>** the `Location`?

- The basic definition of probability remains the same.

- Has the total number of trials changed?
  - YES!



$$p(F = \text{Pitt}|L = \text{Pittsburgh}) = \frac{51300}{68400}$$

<u>Conditional</u> probability of `F` given `L`

Revisit the joint probability

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972}$$

Revisit the joint probability

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972}$$

Multiply by $1 = 68400/68400$:

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972} = \frac{51300}{68400} \cdot \frac{68400}{174972}$$

Revisit the joint probability

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972}$$

Multiply by $1 = 68400/68400$:

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972} = \frac{51300}{68400} \cdot \boxed{\frac{68400}{174972}}$$

$$p(F = \text{Pitt} | L = \text{Pittsburgh}) \cdot \boxed{p(L = \text{Pittsburgh})}$$

Revisit the joint probability

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972}$$

Multiply by $1 = 68400/68400$:

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = \frac{51300}{174972} = \frac{51300}{68400} \cdot \boxed{\frac{68400}{174972}}$$

$$p(L = \text{Pittsburgh} \,\&\, F = \text{Pitt}) = p(F = \text{Pitt} | L = \text{Pittsburgh}) \cdot \boxed{p(L = \text{Pittsburgh})}$$

PRODUCT RULE

Using these rules we can calculate $p(F = \text{Pitt})$

- Start with the sum rule:

$$p(F = \text{Pitt}) = p(L = \text{Pittsburgh}, F = \text{Pitt}) + p(L = \text{State College}, F = \text{Pitt})$$

Using these rules we can calculate $p(F = \text{Pitt})$

- Start with the sum rule:

$$p(F = \text{Pitt}) = p(L = \text{Pittsburgh}, F = \text{Pitt}) + p(L = \text{State College}, F = \text{Pitt})$$

- Substitute in the product rule:

$$p(L = \text{Pittsburgh}, F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

$$p(L = \text{State College}, F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{State College})p(L = \text{State College})$$

Using these rules we can calculate $p(F = \text{Pitt})$

- Start with the sum rule:

$$p(F = \text{Pitt}) = p(L = \text{Pittsburgh}, F = \text{Pitt}) + p(L = \text{State College}, F = \text{Pitt})$$

- Substitute in the product rule:

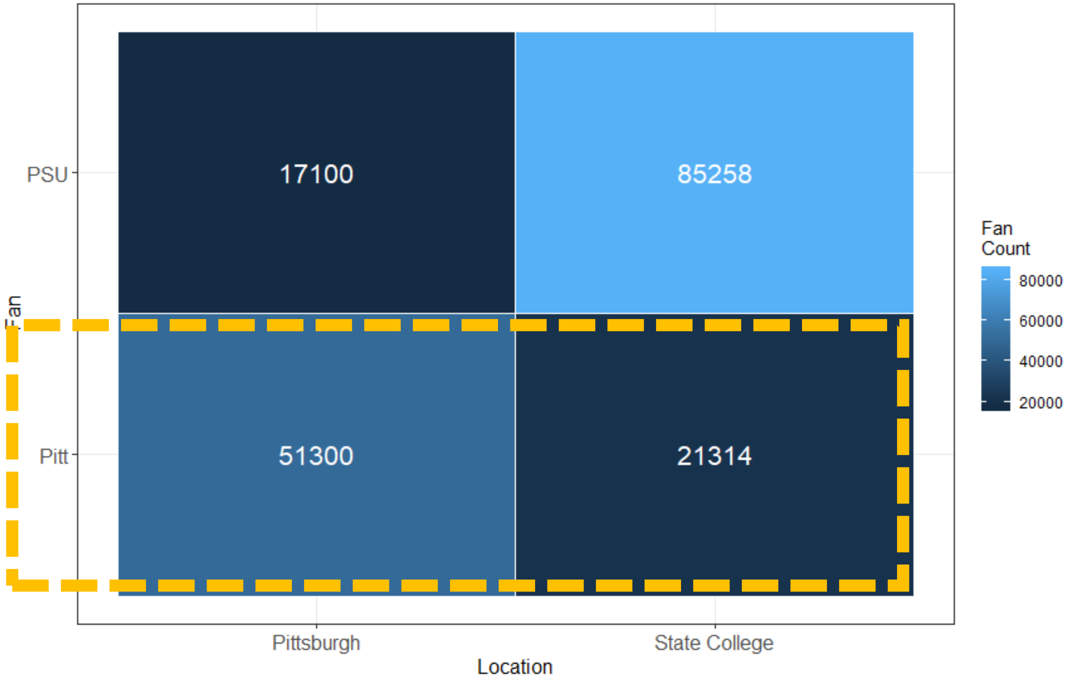
$$p(L = \text{Pittsburgh}, F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

$$p(L = \text{State College}, F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{State College})p(L = \text{State College})$$

- Substitute in the example values:

$$\boxed{p(F = \text{Pitt}) = \frac{51300}{68400} \cdot \frac{68400}{174972} + \frac{21314}{106572} \cdot \frac{106572}{174972} = 0.415}$$

Why is the <u>product rule</u> important?



- In this simple example, we could have calculated $p(F = \text{Pitt})$ using the sum rule alone.
- Follow the same steps we used to calculate $p(L = \text{Pittsburgh})$.
- Sum over the $\mathtt{L}$ variable, for the fixed $F = \text{Pitt}$ value.

In this simple example, we know all quantities, but in a real problem…

- The sum and product rules allow us to rewrite a probability in terms of other probabilities.

- Thus, we can use these rules to write a probability we **DO NOT** know in terms of probabilities we **<u>DO</u>** know!

- Essentially, break a hard problem up into smaller easier pieces to work with.

Let's use this concept to ask a different question: What is the probability $L = \text{Pittsburgh}$, given $F = \text{Pitt}$?

- Start with the joint probability we described previously:

$$p(L = \text{Pittsburgh}, F = \text{Pitt}) = p(F = \text{Pitt} | L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

Let's use this concept to ask a different question: What is the probability $L = \text{Pittsburgh}$, given $F = \text{Pitt}$?

- Start with the joint probability we described previously:

$$p(L = \text{Pittsburgh}, F = \text{Pitt}) = p(F = \text{Pitt} | L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

- We can also write:

$$p(F = \text{Pitt}, L = \text{Pittsburgh}) = p(L = \text{Pittsburgh} | F = \text{Pitt})p(F = \text{Pitt})$$

Let's use this concept to ask a different question: What is the probability $L = \text{Pittsburgh}$, given $F = \text{Pitt}$?

- Start with the joint probability we described previously:

$$p(L = \text{Pittsburgh}, F = \text{Pitt}) = p(F = \text{Pitt} | L = \text{Pittsburgh}) p(L = \text{Pittsburgh})$$

- We can also write:

$$p(F = \text{Pitt}, L = \text{Pittsburgh}) = p(L = \text{Pittsburgh} | F = \text{Pitt}) p(F = \text{Pitt})$$

- Use the **symmetry property**:

$$p(L = \text{Pittsburgh}, F = \text{Pitt}) = p(F = \text{Pitt}, L = \text{Pittsburgh})$$

Substitute in the product rule relationships into each side of the symmetry property

$$p(L = \text{Pittsburgh}|F = \text{Pitt})p(F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

Substitute in the product rule relationships into each side of the symmetry property

$$p(L = \text{Pittsburgh}|F = \text{Pitt})p(F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

- Rearrange to write the probability of interest in terms of probabilities we already know:

$$p(L = \text{Pittsburgh}|F = \text{Pitt}) = \frac{p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})}{p(F = \text{Pitt})}$$

Substitute in the product rule relationships into each side of the symmetry property

$$p(L = \text{Pittsburgh}|F = \text{Pitt})p(F = \text{Pitt}) = p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})$$

- Rearrange to write the probability of interest in terms of probabilities we already know:

$$p(L = \text{Pittsburgh}|F = \text{Pitt}) = \frac{p(F = \text{Pitt}|L = \text{Pittsburgh})p(L = \text{Pittsburgh})}{p(F = \text{Pitt})}$$

The above relationship is **BAYES' THEOREM**!!!!

Evaluate using the example values

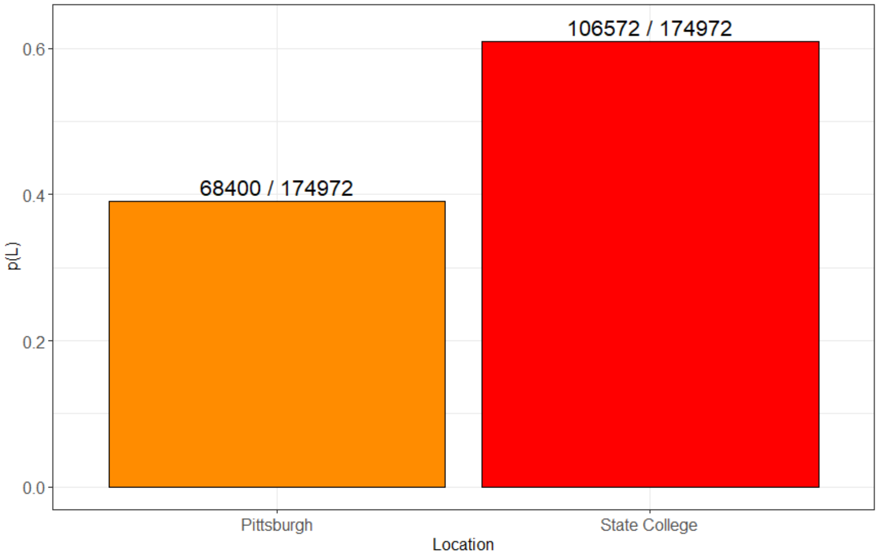
$$p(L = \text{Pittsburgh}|F = \text{Pitt}) = \frac{\frac{51300}{68400} \cdot \frac{68400}{174972}}{0.415} = 0.706$$

But, what does $p(L = \text{Pittsburgh}|F = \text{Pitt})$ represent?

- Consists of three terms.

But, what does $p(L = \text{Pittsburgh}|F = \text{Pitt})$ represent?

Initially, we know the counts at `Pittsburgh` and `State College`.

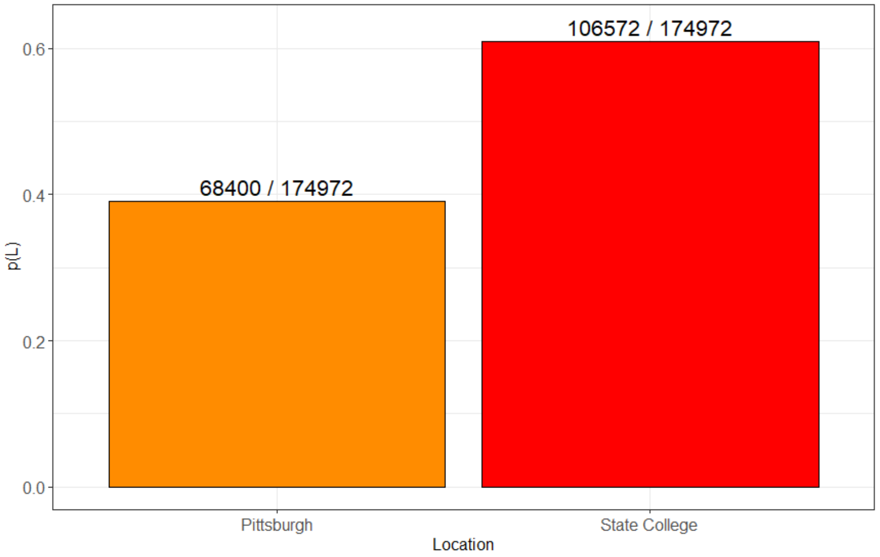


$$p(L = \text{Pittsburgh}) = \frac{68400}{174972} = 0.391$$

$p(L = \text{Pittsburgh})$ is the **<u>PRIOR</u>** probability, ***before*** saying anything about the `Fan` type.

But, what does $p(L = \text{Pittsburgh}|F = \text{Pitt})$ represent?

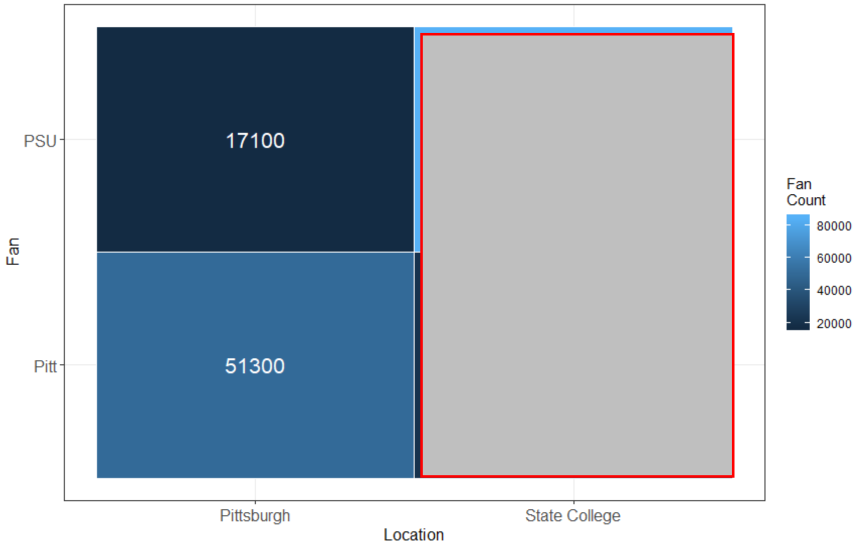
Initially, we know the counts at `Pittsburgh` and `State College`.



$$p(L = \text{Pittsburgh}) = \frac{68400}{174972} = 0.391$$

$p(L = \text{Pittsburgh})$ is the **PRIOR** probability, ***before*** saying anything about the `Fan` type.

Fixing $L = \text{Pittsburgh}$ changes the total number of trials!

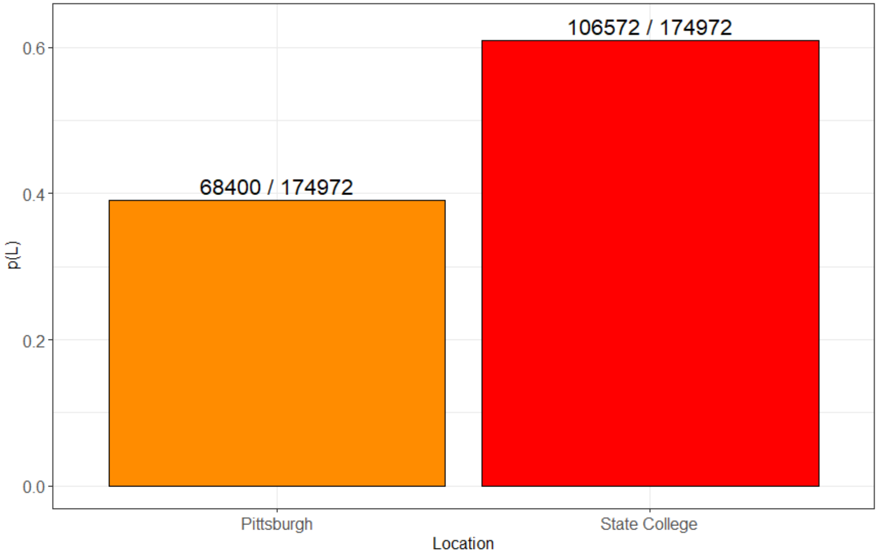


Total number of trials is now 68400

$p(F = \text{Pitt}|L = \text{Pittsburgh})$ is called the **LIKELIHOOD**.

But, what does $p(L = \text{Pittsburgh}|F = \text{Pitt})$ represent?

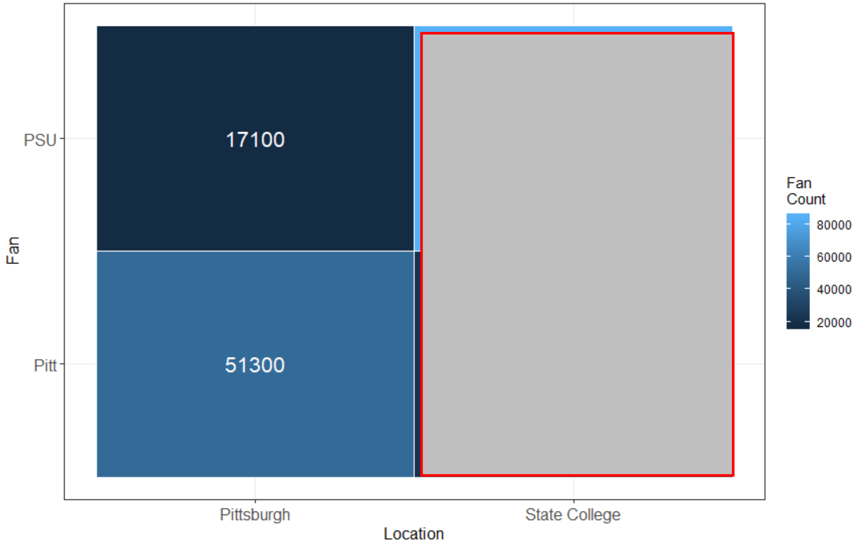
Initially, we know the counts at `Pittsburgh` and `State College`.



$$p(L = \text{Pittsburgh}) = \frac{68400}{174972} = 0.391$$

$p(L = \text{Pittsburgh})$ is the **PRIOR** probability, ***before*** saying anything about the `Fan` type.

Fixing $L = \text{Pittsburgh}$ changes the total number of trials!



Total number of trials is now 68400

$p(F = \text{Pitt}|L = \text{Pittsburgh})$ is called the **LIKELIHOOD**.

$p(F = \text{Pitt})$ is the **average** probability of a Pitt fan.

Compute using the sum and product rule, as we did previously.

Serves as the normalizing constant that guarantees the probability is between 0 and 1.

$p(F = \text{Pitt})$ referred to as the **EVIDENCE** or **MARGINAL LIKELIHOOD**.

Combining the PRIOR, LIKELIHOOD, and EVIDENCE yields the <u>POSTERIOR</u>

$$\mathrm{Posterior} = \frac{\mathrm{Likelihood} \times \mathrm{Prior}}{\mathrm{Evidence}}$$

We updated our prior ***belief*** given the observations!

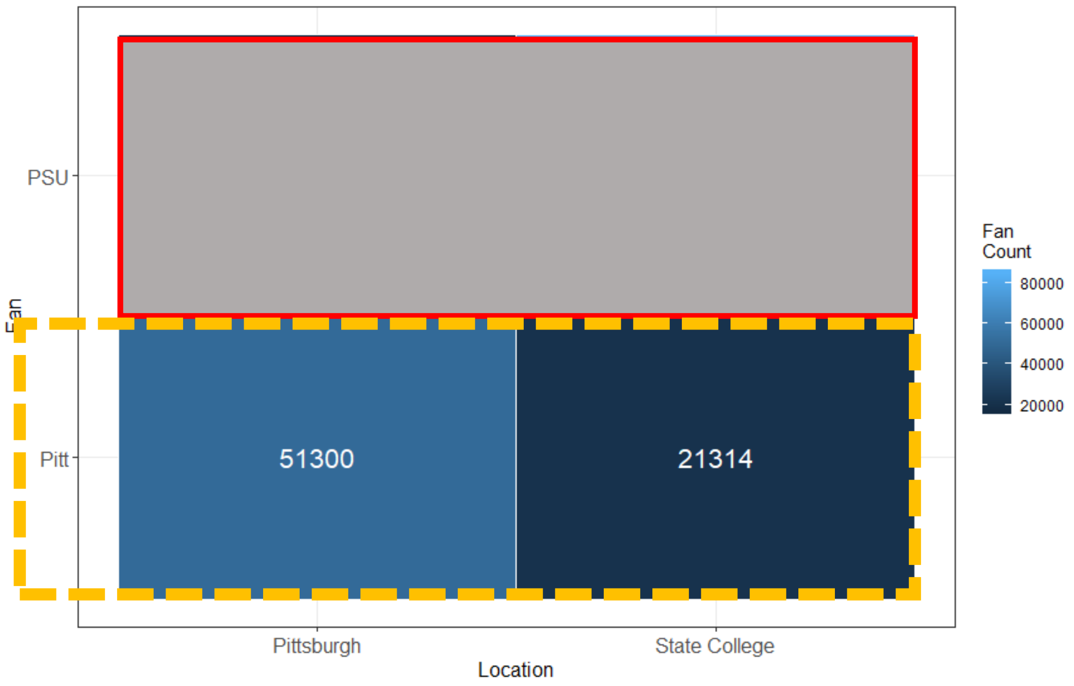
The calculations and terminology may seem a little confusing, but we already knew how to calculate $p(L = \text{Pittsburgh}|F = \text{Pitt})$...

- Follow the same conditioning procedure we saw earlier.

- Fixing $F = \text{Pitt}$ changes the total number of trials to 72614.

- The conditional probability is:

$$\frac{51300}{72614} = 0.706$$



In many respects, probability is just counting!

- Get the bookkeeping correct, and the rest basically takes care of itself.

- In this course, we will use a lot of math.

- If the math appears overwhelming, just remember we're just counting!

Probability distributions

In our college football example, we were given counts for each of the 4 combinations.

- Beaver Stadium at State College has a capacity 106572 fans.
- We won't be able to ask everyone there!
- How can we represent the population attending the game?

In our college football example, we were given counts for each of the 4 combinations.

- Beaver Stadium at State College has a capacity 106572 fans.

- We won’t be able to ask everyone there!

- How can we represent the population attending the game?
  - SAMPLING!!!!

Assumptions

- Maintain the assumption that all fans are independent.

- Assume our sample size is very small relative to the population size.

Notation

- Instead of denoting `Fan` as `F`, we will now use the general variable x.

- Encode `Fan=Pitt` as $x = 1$ and `Fan=PSU` as $x = 0$.

The probability of `Fan` = `Pitt`, $x = 1$, will be denoted by the parameter μ

- The above statement can be written as:

$$p(x = 1|\mu) = \mu$$

The probability of `Fan = Pitt`, $x = 1$, will be denoted by the parameter μ

- The above statement can be written as:

$$p(x = 1|\mu) = \mu$$

- Since we have either $x = 1$ or $x = 0$, the probability that $x = 0$ is:

$$p(x = 0|\mu) = 1 - \mu$$

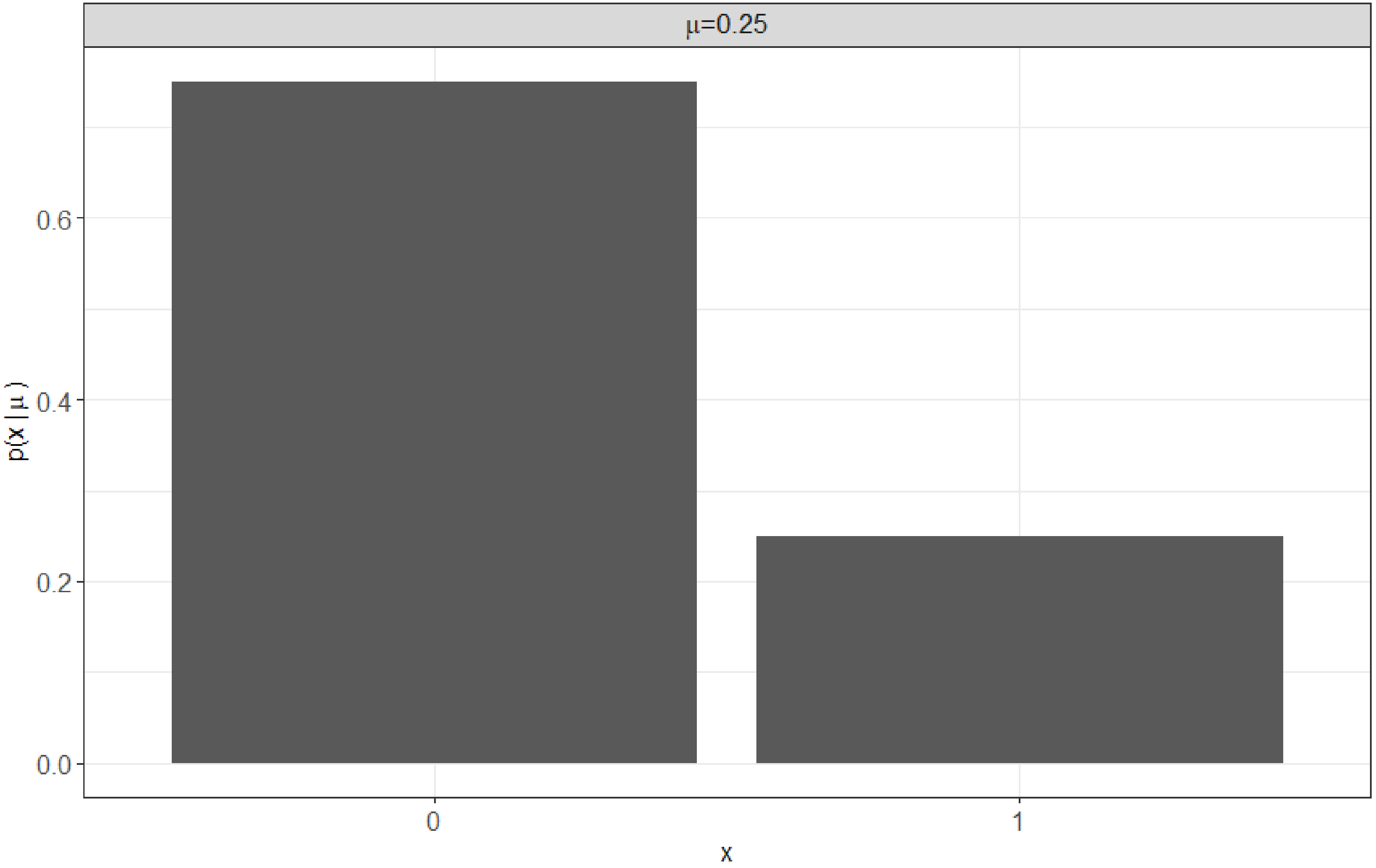
The **probability mass function** over possible outcomes can be more compactly written as:

$$\boxed{p(x|\mu) = \mathrm{Bernoulli}(x|\mu) = \mu^x (1-\mu)^{1-x}}$$

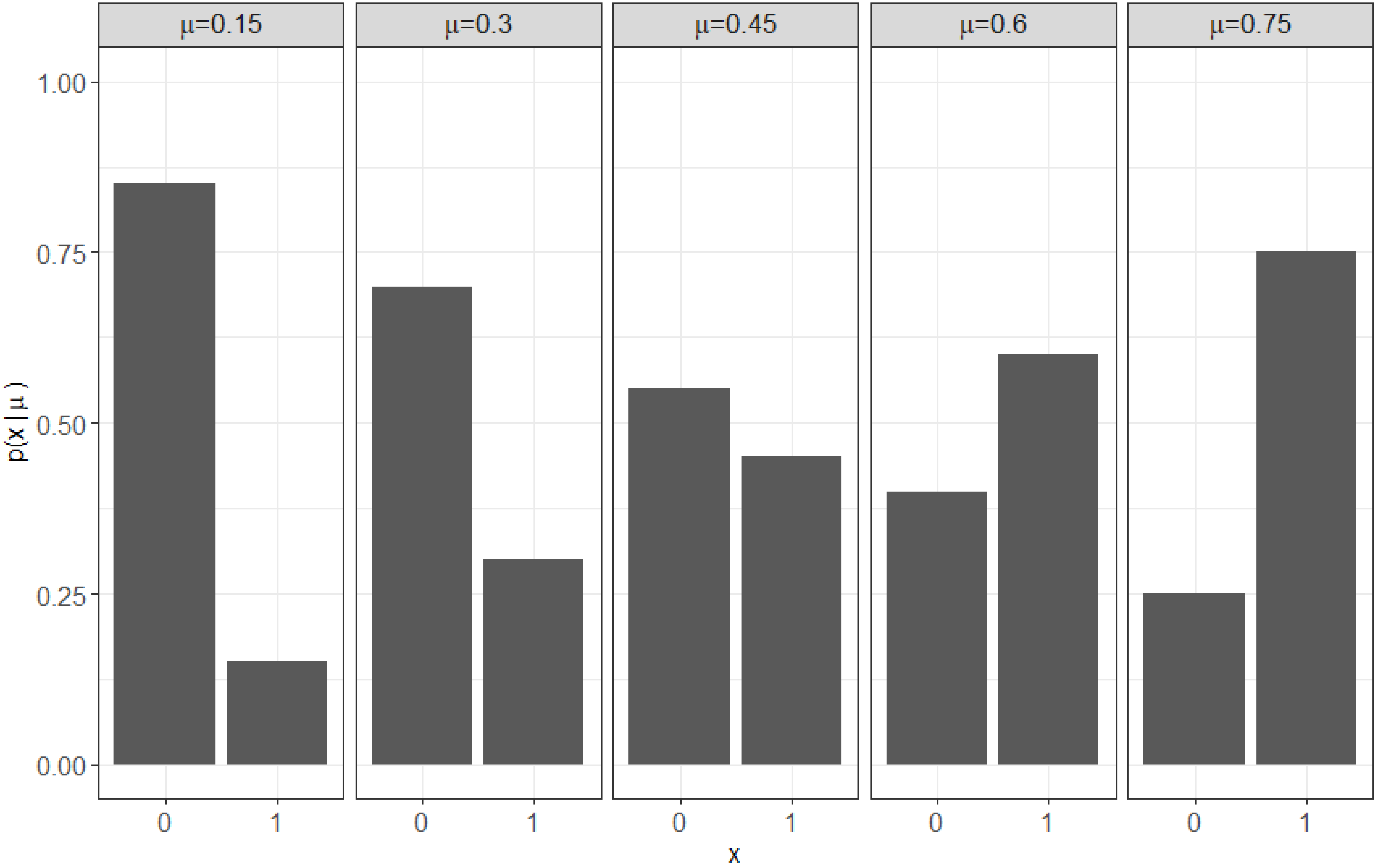
Referred to as the Bernoulli distribution after Jacob Bernoulli

Note: μ is a probability and so is bounded: $0 \leq \mu \leq 1$

The Bernoulli distribution PMF for a single μ



Bernoulli PMF for multiple μ values



The Bernoulli distribution can be used to represent many different real-life applications

- Today we introduced the concept in terms of college football fans.

- Typically, the Bernoulli distribution is introduced with the canonical **coin flip** example.

- Applicable in sports, engineering, manufacturing, medicine, marketing, etc…

The Bernoulli distribution can be used to represent many different real-life applications

- Today we introduced the concept in terms of college football fans.

- Typically, the Bernoulli distribution is introduced with the canonical **coin flip** example.

- Applicable in sports, engineering, manufacturing, medicine, marketing, etc…

- The Bernoulli distribution is applicable for many **BINARY OUTCOME** situations.

Back to our college football example…

- Suppose we ask 4 random people which team they are cheering for, and we get the following answers:

Person	Fan
1	PSU
2	PSU
3	Pitt
4	PSU

In terms of the **encoded** variable x:

Person	Fan	x
1	PSU	0
2	PSU	0
3	Pitt	1
4	PSU	0

What is the probability of the sequence we observed?

- Start with the joint distribution:

$$p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0|\mu)$$

- Assuming each person is <u>**independent**</u> we can factorize the joint distribution into:

$$p(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0|\mu) = p(x_1 = 0)p(x_2 = 0)\, p(x_3 = 1)\, p(x_4 = 0)$$

Drop μ on the right-hand side (RHS) to save space

Substitute in the expression from the Bernoulli PMF for each person

$$p(x_1 = 0)p(x_2 = 0)p(x_3 = 1)p(x_4 = 0) = (1 - \mu)(1 - \mu)\mu(1 - \mu)$$

Drop μ to save space

Now generalize from asking 4 people, to asking N people

- Denote the sequence of responses as: $\mathbf{x} = \{x_1, x_2, \dots, x_n, \dots, x_N\}$

- Assuming each person is **independent** we can factorize the joint distribution just as we did before:

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} \{\mu^{x_n}(1-\mu)^{1-x_n}\} = \prod_{n=1}^{N} \mathrm{Bernoulli}(x_n|\mu)$$

Now generalize from asking 4 people, to asking N people

- Denote the sequence of responses as: $\mathbf{x} = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$

- Assuming each person is <u>**independent**</u> we can factorize the joint distribution just as we did before:

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} \{\mu^{x_n}(1-\mu)^{1-x_n}\} = \prod_{n=1}^{N} \mathrm{Bernoulli}(x_n|\mu)$$

Likelihood Function!!!

What happens when μ is unknown?

- If we have the observed sequence $\mathbf{x} = \{x_1, x_2, \dots, x_n, \dots, x_N\}$ we can estimate μ from the data.

- Which value of μ is best?

What happens when μ is unknown?

- If we have the observed sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_n, \ldots, x_N\}$ we can estimate μ from the data.

- Which value of μ is best?

- The value that **MAXIMIZES** the likelihood!

$$\hat{\mu} = \mu_{ML} = \underset{\mu \in [0,1]}{\operatorname{argmax}}\, p(\mathbf{x}|\mu)$$

Rather than working with the likelihood directly…we will maximize the *log-likelihood*

- The log-likelihood for the N independent observations:

$$\log[p(\mathbf{x}|\mu)] = \sum_{n=1}^{N}\{x_n \log[\mu] + (1 - x_n)\log[1 - \mu]\}$$

Rearrange the log-likelihood

$$\log[p(\mathbf{x}|\mu)] = \log[\mu] \sum_{n=1}^{N} \{x_n\} + \log[1-\mu] \sum_{n=1}^{N} \{1 - x_n\}$$

Rearrange the log-likelihood

$$\log[p(\mathbf{x}|\mu)] = \log[\mu] \underbrace{\sum_{n=1}^{N} \{x_n\}} + \log[1-\mu] \underbrace{\sum_{n=1}^{N} \{1 - x_n\}}$$

Number of Pitt fans, or more generally number of events M.

Number of PSU fans, or more generally number of times we **did not** observe the event $N - M$.

Rearrange the log-likelihood

$$\boxed{\log[p(\mathbf{x}|\mu)] = \log[\mu] \cdot M + \log[1-\mu] \cdot (N - M)}$$

To optimize, calculate the derivative of the log-likelihood with respect to $\boldsymbol{\mu}$

$$\frac{\partial}{\partial \mu}\{\log[p(\mathbf{x}|\mu)]\} = \frac{\partial}{\partial \mu}\{\log[\mu] \cdot M\} + \frac{\partial}{\partial \mu}\{\log[1-\mu] \cdot (N-M)\}$$

To optimize, calculate the derivative of the log-likelihood with respect to $\boldsymbol{\mu}$

$$\frac{\partial}{\partial \mu}\{\log[p(\mathbf{x}|\mu)]\} = \underbrace{\frac{\partial}{\partial \mu}\{\log[\mu] \cdot M\}}_{\frac{M}{\mu}} + \underbrace{\frac{\partial}{\partial \mu}\{\log[1-\mu] \cdot (N-M)\}}_{\frac{-(N-M)}{1-\mu}}$$

Set the derivative equal to zero and solve for μ_{ML}

$$\frac{\partial}{\partial \mu}\{\log[p(\mathbf{x}|\mu)]\} = 0 = \frac{M}{\mu_{ML}} - \frac{N-M}{1-\mu_{ML}}$$

Set the derivative equal to zero and solve for μ_{ML}

$$\frac{\partial}{\partial \mu}\{\log[p(\mathbf{x}|\mu)]\} = 0 = \frac{M}{\mu_{ML}} - \frac{N-M}{1-\mu_{ML}}$$

$$\Downarrow$$

$$\frac{(1-\mu_{ML})\cdot M - \mu_{ML}\cdot(N-M)}{\mu_{ML}\cdot(1-\mu_{ML})} = 0$$

Set the derivative equal to zero and solve for μ_{ML}

$$\frac{\partial}{\partial \mu}\{\log[p(\mathbf{x}|\mu)]\} = 0 = \frac{M}{\mu_{ML}} - \frac{N-M}{1-\mu_{ML}}$$

$$\Downarrow$$

$$\frac{(1-\mu_{ML}) \cdot M - \mu_{ML} \cdot (N-M)}{\mu_{ML} \cdot (1-\mu_{ML})} = 0$$

$$\Downarrow$$

$$M - \mu_{ML}M - \mu_{ML}N + \mu_{ML}M = 0 \rightarrow M - \mu_{ML}N = 0$$

The maximum likelihood estimate (MLE) for μ is just based on counting!

$$\boxed{\mu_{ML} = \frac{M}{N} = \frac{1}{N}\sum_{n=1}^{N}\{x_n\}}$$