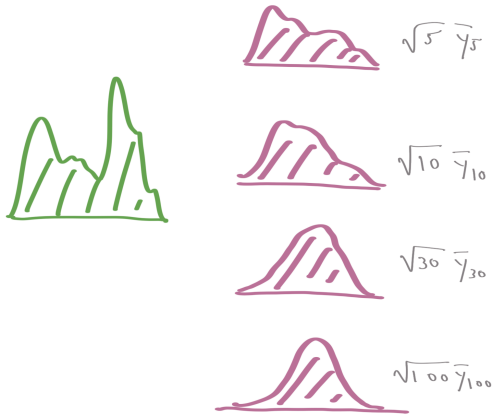
Probability Review



Statistical Experimental Design

Kris Sankaran | UW Madison | 14 September 2021

Sample $\rightarrow$ Population Properties

- The distribution $\mathbf{P}$ summarizes our model of the world
- We will be happy if we can make precise statements about it
  - Where is the center?
  - How spread out is it?
  - What is its shape?
  - How many peaks does it have?
  - ...

Statistical Estimators

- We only have access to a sample $x_1, \ldots, x_n$ from $\mathbf{P}$.
  - Assume they are all independent replicates
- We can define functions of the sample in order to estimate properties of $\mathbf{P}$
  - $\bar{x} \approx \mu\left(\mathbf{P}\right)$
  - $\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \approx \sigma^2\left(\mathbf{P}\right)$

How to evaluate estimators?

- Suppose we have 100 samples from a random normal distribution with unknown mean μ
- Which is a better estimator of the mean? Why?

$$\text{Option 1: } \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

$$\text{Option 2: } \bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i$$

Low Bias and Variance

The code below generates 5000 datasets with population means 2.5, then computes the two estimators.

```
n_sim <- 5e3
datasets <- matrix(rnorm(n_sim * 100, 2.5), n_sim, 100)
means <- data.frame(
  id = 1:n_sim,
  partial = rowMeans(datasets[, 1:10]),
  full = rowMeans(datasets)
)
```

Low Bias and Variance

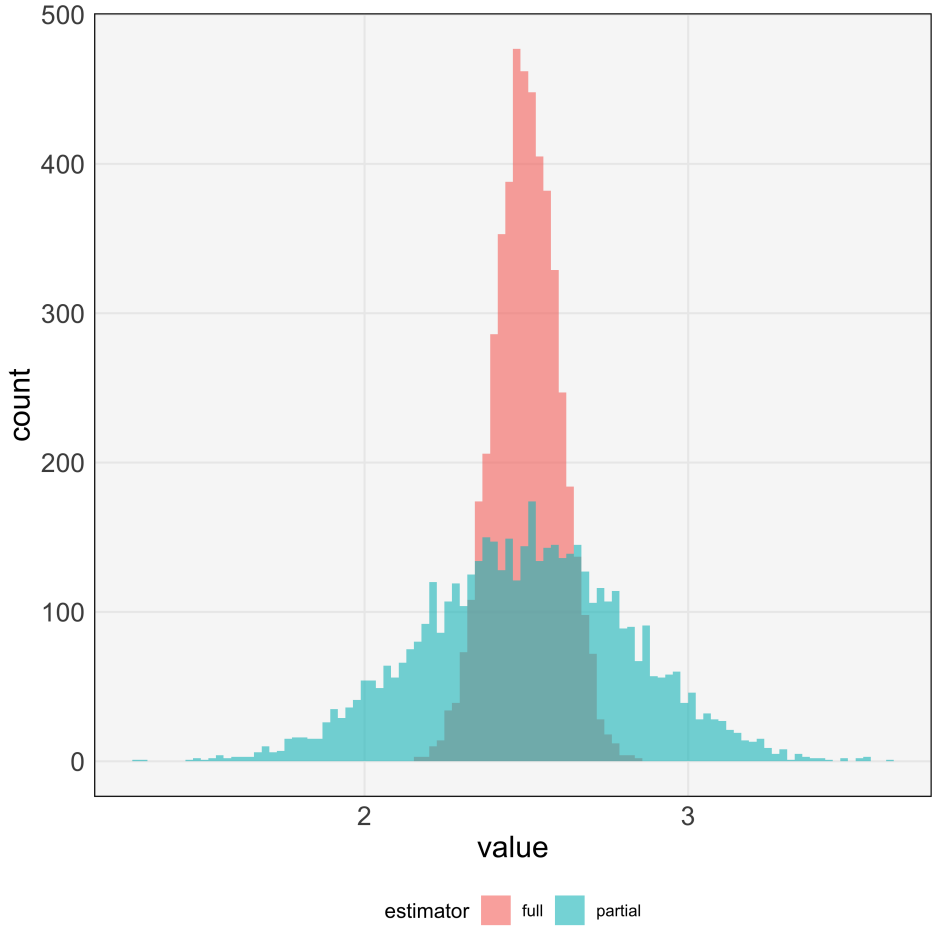
The code below generates 5000 datasets with population means 2.5, then computes the two estimators.

```
head(means)
```

```
##   id  partial     full
## 1  1 2.291631 2.450778
## 2  2 2.864438 2.627423
## 3  3 2.207296 2.504182
## 4  4 2.610049 2.636441
## 5  5 2.148647 2.391529
## 6  6 2.178199 2.409121
```

Low Bias and Variance

- Unbiased: The statistic is centered around the truth
- Low Variance: The spread of the statistic is low
- Using all the data gives an estimate with lower variance than using only a fraction



Discussion (if time)

- Can you come up with two estimates of the population mean, one of which is biased but low variance, and the other which has low / no bias and high variance?
- Can you design a simulation to compare the bias and variance of the following estimates of the population standard deviation?

$$\text{Option 1: } \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$\text{Option 2: } 1.483\left[\frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}_i\right|\right]$$

Central Limit Theorem

Theorem Statement

If y_i are drawn i.i.d. from some distribution with mean μ and variance σ^2, then

$$\frac{\sqrt{n}\left(\bar{y}-\mu\right)}{\sigma} \rightarrow \mathcal{N}\left(0,1\right).$$

Theorem Importance

- This theorem reduces calculations across arbitrary distributions into calculations with normal distributions.
- It helps in finding approximate reference distributions without having to resort to simulation

Exercise Warm-Up

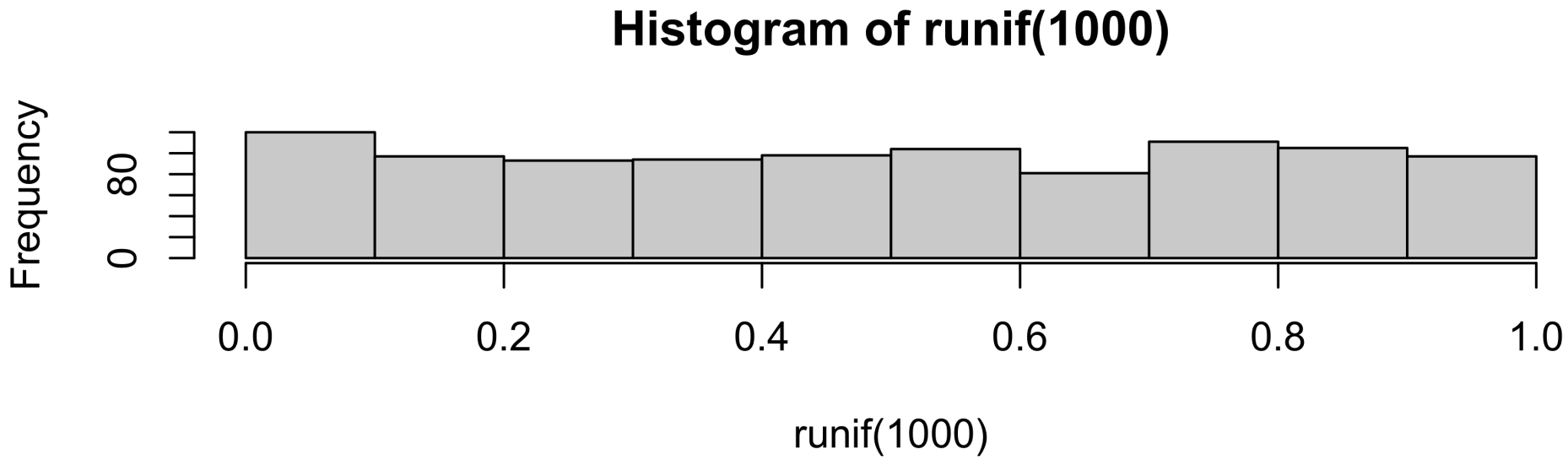
What will be the shape of the histogram in the block below? Why?

```
#hist(runif(1000))
```

Exercise Warm-Up

What will be the shape of the histogram in the block below? Why?

```
hist(runif(1000))
```



Exercise Warm-Up

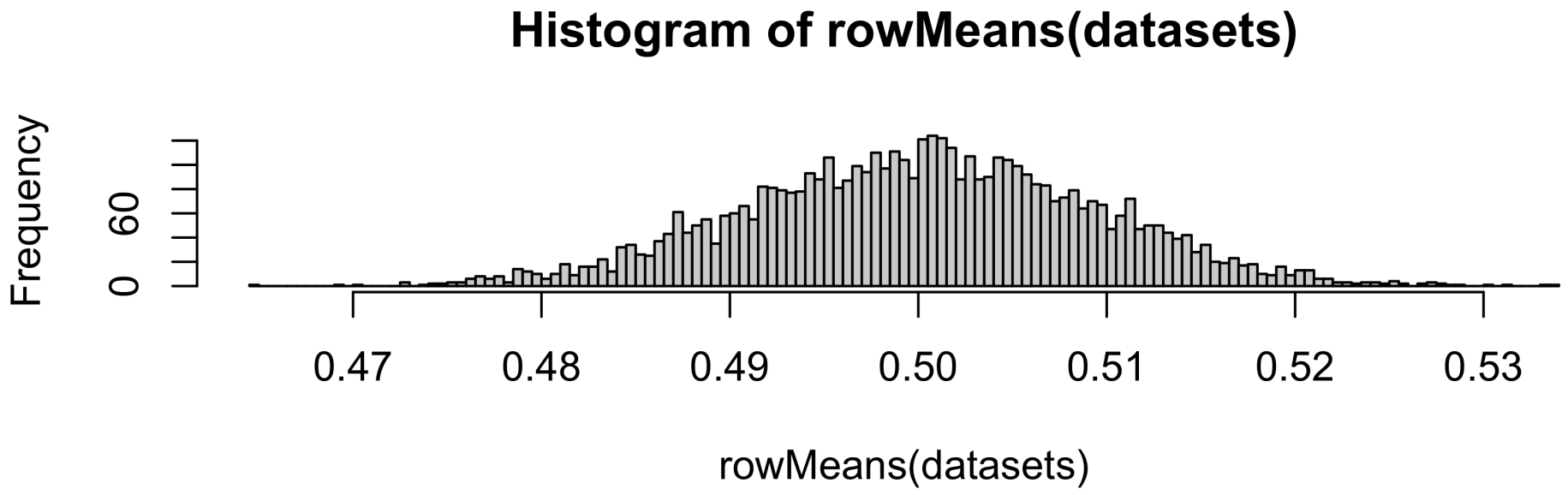
What will be the shape of the histogram in the block below? Why?

```
datasets <- matrix(runif(n_sim * 1000), nrow = n_sim, ncol = 1000)
#hist(rowMeans(datasets), breaks = 100)
```

Exercise Warm-Up

What will be the shape of the histogram in the block below? Why?

```
datasets <- matrix(runif(n_sim * 1000), nrow = n_sim, ncol = 1000)
hist(rowMeans(datasets), breaks = 100)
```



Exercise

What will be the differences between the three histograms below? Why?

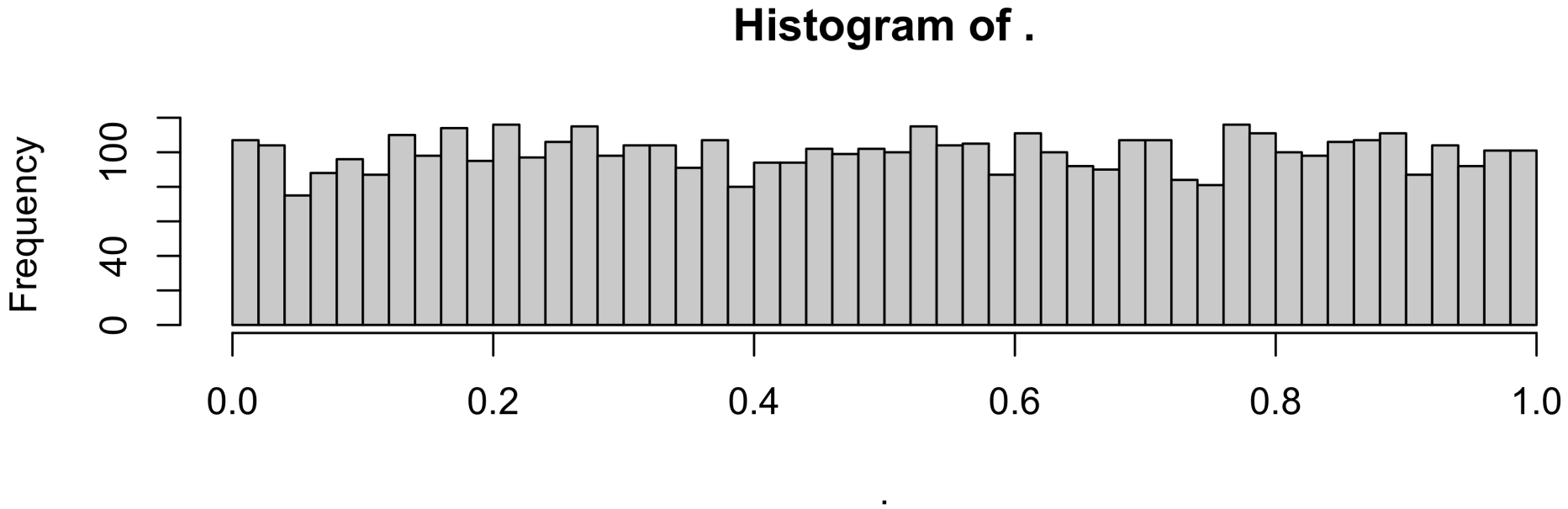
```
plot_hist <- function(sample_size) {
  matrix(runif(n_sim * sample_size), n_sim, sample_size) %>%
    rowMeans() %>%
    hist(breaks = 50)
}

#plot_hist(1)
#plot_hist(2)
#plot_hist(1000)
```

Exercise

What will be the differences between the three histograms below? Why?

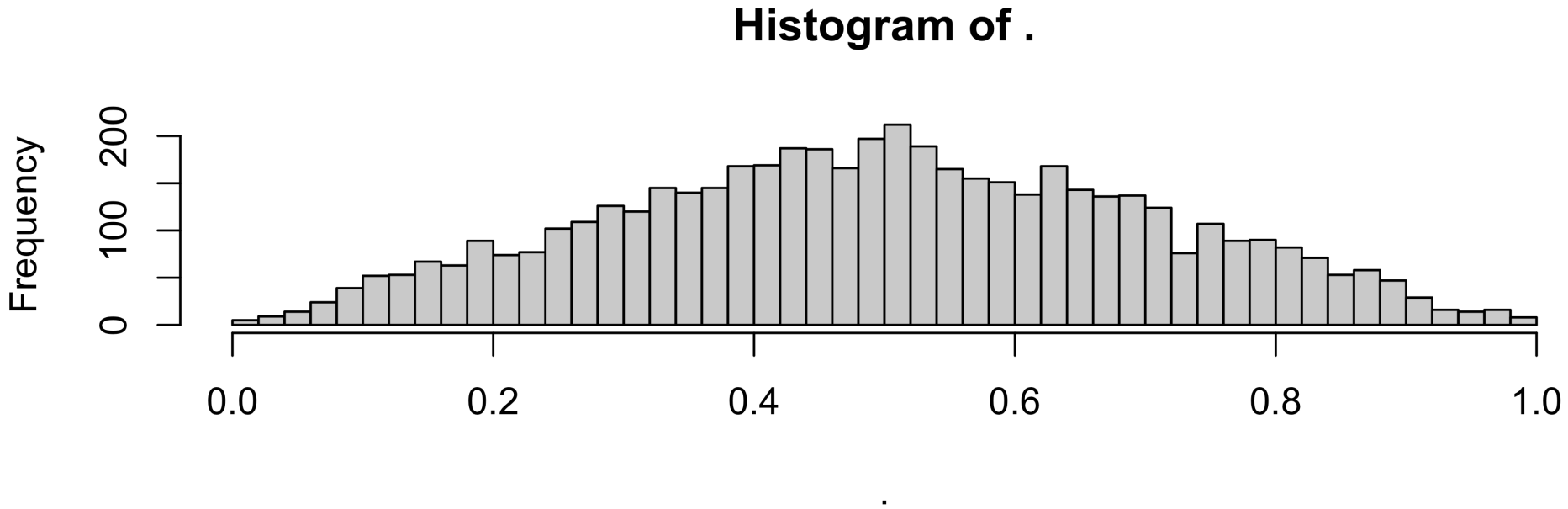
```
plot_hist(1)
```



Exercise

What will be the differences between the three histograms below? Why?

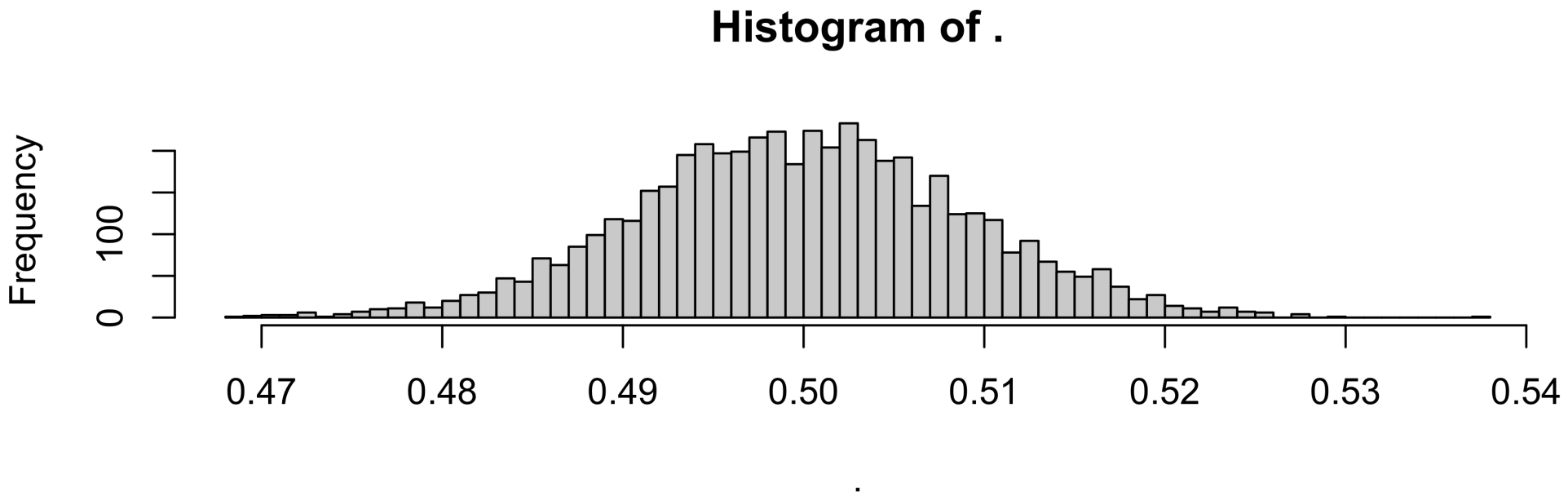
```
plot_hist(2)
```



Exercise

What will be the differences between the three histograms below? Why?

```
plot_hist(1000)
```



Useful Distributions

R Syntax

- `r{name of density}(n)` will sample n points
- `d{name of density}(x)` will compute the probability density at at x
- `p{name of density}(x)` will integrate the density up to x
- `q{name of density}(p)` will find the x value of the density at the p quantile

t Distribution

- We can use the `dt` function to compute the density of the t distribution.
- Evaluate over a grid of `x` values to make a plot

```
x <- seq(-3, 3, length.out = 100)
data.frame(x, density = dt(x, df=2)) %>%
  ggplot() +
  geom_line(aes(x, density))
```

t Distribution

It has a hyperparameter, called the "degrees-of-freedom" (df). Smaller df means heavier tails.

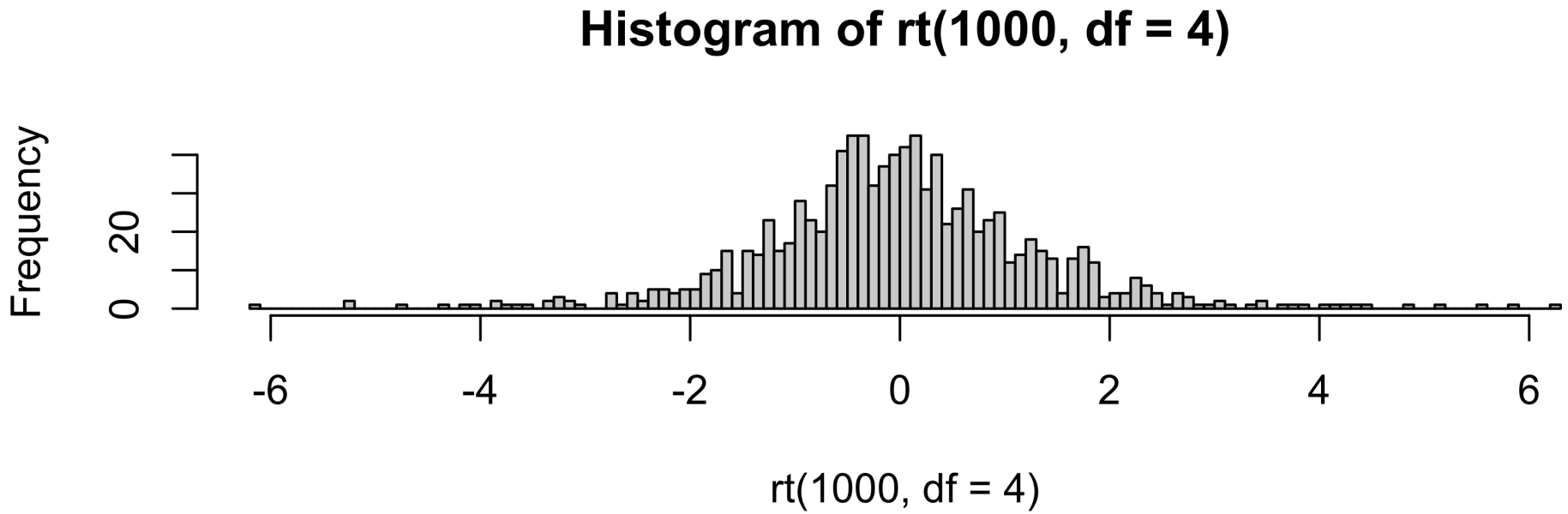
```
densities <- map_dfr(
  seq(1, 10, .5),
  ~ data.frame(x, density = dt(x, .), df = .)
)

ggplot(densities) +
  geom_line(aes(x, density, col = df, group = df)) +
  theme(legend.position = "right")
```

t Distribution

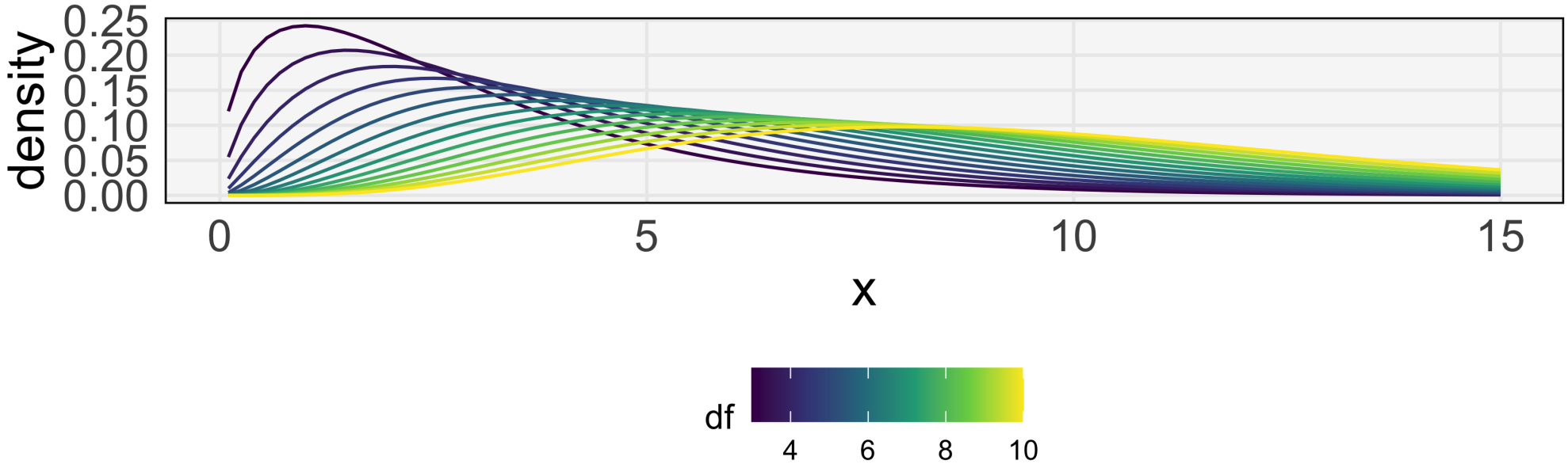
Here are examples generating samples and computing quantiles.

```
hist(rt(1000, df = 4), breaks = 100)
```



chi-square Distribution

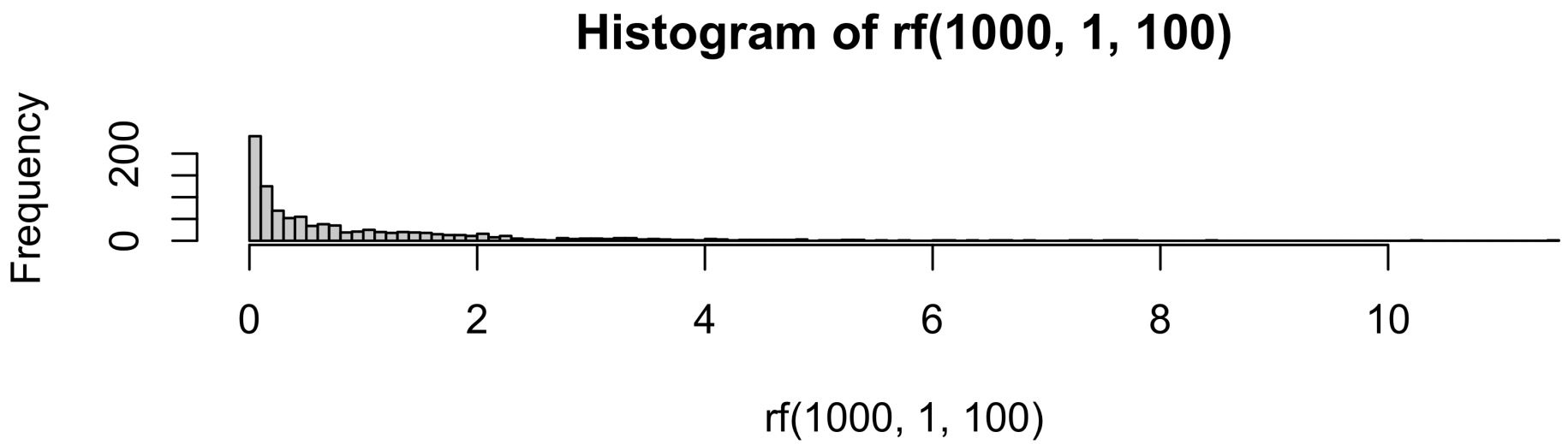
The chi-square distribution is nonnegative with one parameter and can be referenced using `(prefix)chisq`.



F Distribution

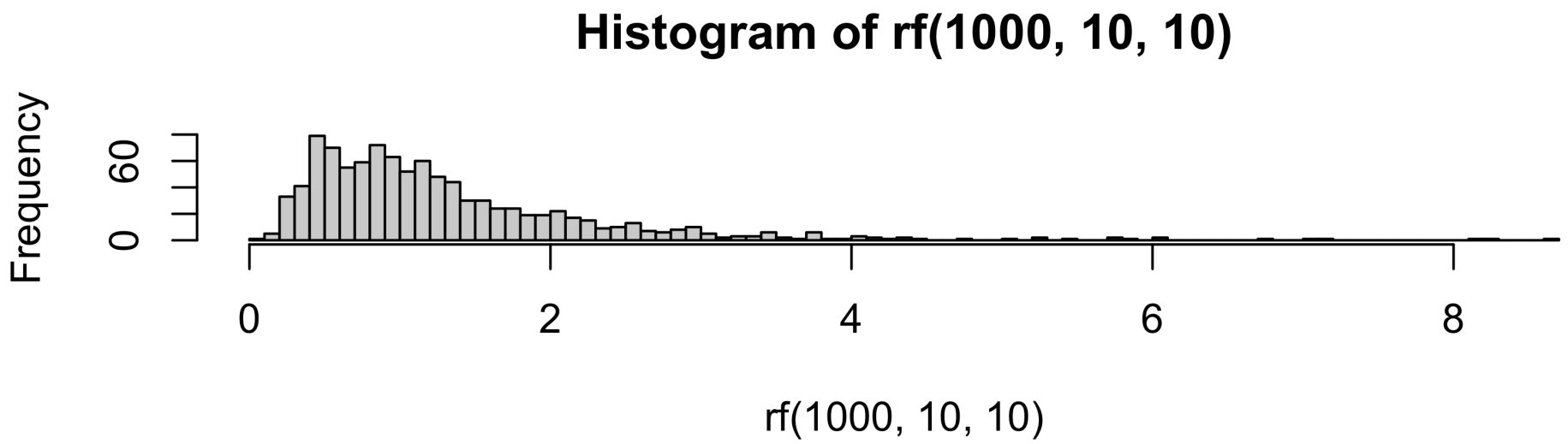
The F distribution is also nonnegative, but has two parameters.

```
hist(rf(1000, 1, 100), breaks = 100)
```



F Distribution

```
hist(rf(1000, 10, 10), breaks = 100)
```



Where these distributions arise

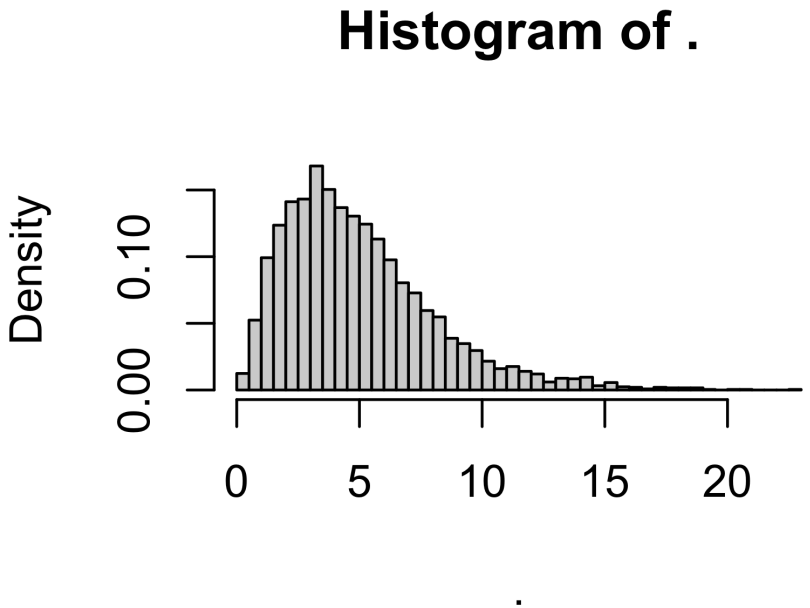
Main Idea

- We will often want the distribution of a particular statistic
- We may know the distribution of individual terms within the statistic
- Learning how one distribution arises as a function of another is key

Chi-square Distribution

This distribution arises as the sum-of-squares of standard normals. If $z_k \sim \mathcal{N}(0,1)$, then $\sum_{k=1}^{K} z_k^2 \sim \chi^2_K$.

```
rchisq(n_sim, 5) %>%
  hist(breaks = 50, freq=F, ylim = c(0, .18
```

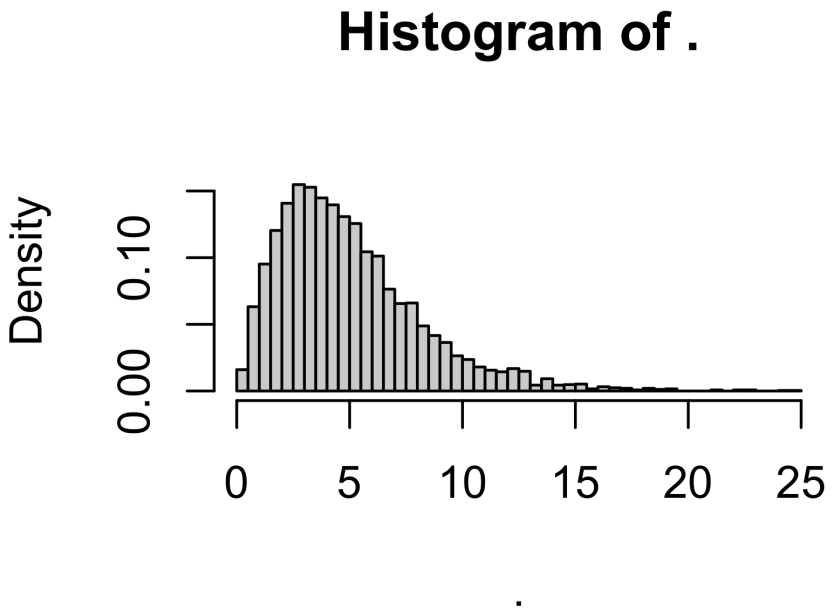


.

Chi-square Distribution

This distribution arises as the sum-of-squares of standard normals. If $z_k \sim \mathcal{N}(0,1)$, then $\sum_{k=1}^{K} z_k^2 \sim \chi^2_K$.

```
matrix(rnorm(n_sim * 5)^2, n_sim, 5) %>%
  rowSums() %>%
  hist(breaks = 50, freq=F, ylim = c(0, .18
```



.

Chi-square Distribution

A related (but nontrivial) fact is that if $y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$,

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \sim \chi^2_{n-1}$$

Chi-square Distribution

```
hist(rchisq(n_sim, 9), breaks = 50, col = rgb(0, 0, 1, .6))
datasets <- rerun(n_sim, rnorm(10, 2.5, 1))
ss <- map_dbl(datasets, ~ sum((. - mean(.)) ^ 2))
hist(ss, breaks = 50, col = rgb(0, 1, 0, 0.6), add = TRUE)
```

t Distribution

The t distribution can be formed as the ratio,

$$\frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_K}{K}}}$$

This ratio often occurs when we standardize using an estimate of the standard deviation,

$$\frac{\sqrt{n}(\bar{y}-\mu)}{S}$$

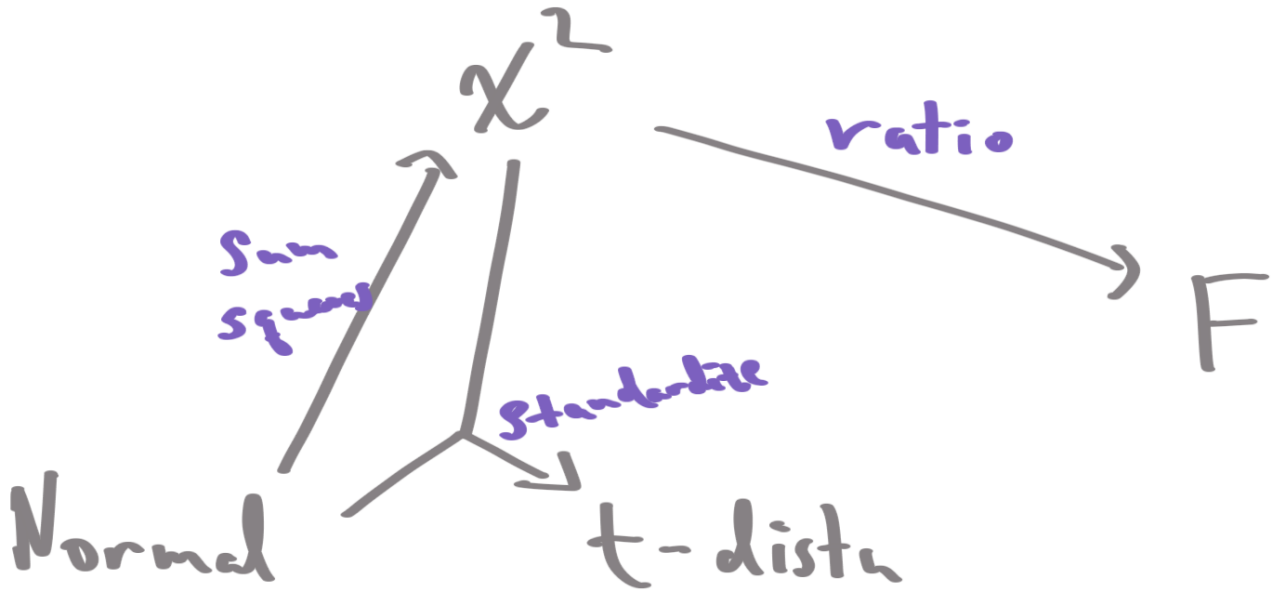
F Distribution

This distribution arises as the ratio,

$$F_{u,v} = \frac{\frac{1}{u}\chi^2_u}{\frac{1}{v}\chi^2_v}.$$

Since chi-squares come up whenever we compute sums-of-squares of normals, this statistic will arise whenever we want to compare two different sums-of-squares.

Summary



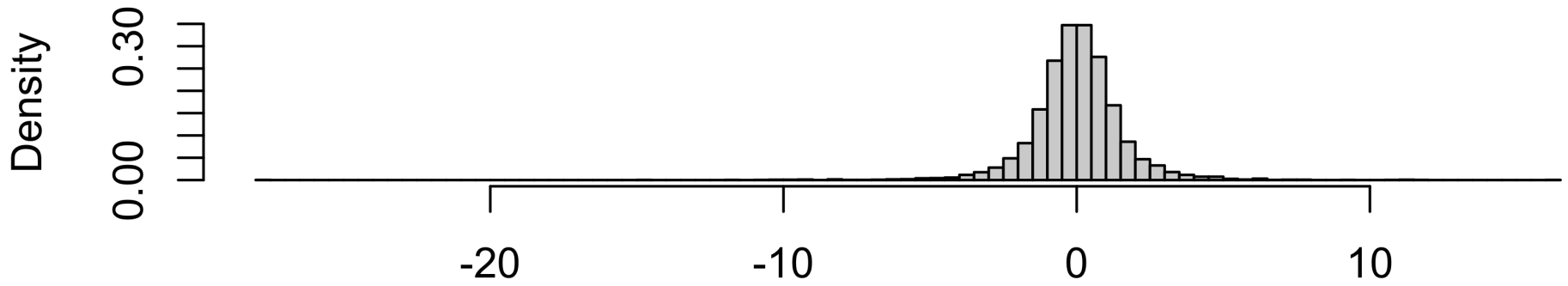
Exercise

Write a small simulation to generate t-distributed variables (with whatever d.f. you want), without using `rt`.

- Bonus: Make a histogram
- Bonus: Use only `rnorm()`.

Solution 1

```
df <- 3
(rnorm(n_sim) / sqrt(rchisq(n_sim, df) / df)) %>%
  hist(breaks = 100, freq = F, xlab = NULL, main = NULL)
```



Solution 2

```
normalized_stat <- function(n) {
  y <- rnorm(n)
  sqrt(n) * mean(y) / sd(y)
}

unlist(rerun(n_sim, normalized_stat(df + 1))) %>%
  hist(breaks = 100, freq = F, xlab = NULL, main = NULL)
```