STATS 2107
Statistical Modelling and Inference II

Workshop 4: Sampling distributions part 1

Matt Ryan

School of Mathematical Sciences, University of Adelaide

Semester 2 2022

The sampling distribution of the sample mean

What is a sampling distribution?

Suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample, and T is a statistic on the Y_i. Then the distribution of T is called the *sampling distribution*.

The sample mean

For example, suppose each $Y_i \sim N(\mu, \sigma^2)$ and $T = \bar{Y}$. Then the sampling distribution is

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

What is meant by sampling distribution?

$$
\begin{array}{cccccc}
Y_{11}, & Y_{12}, & \ldots, & Y_{1n} & \rightarrow & T_1 \\
Y_{21}, & Y_{22}, & \ldots, & Y_{2n} & \rightarrow & T_2 \\
Y_{31}, & Y_{32}, & \ldots, & Y_{3n} & \rightarrow & T_3 \\
Y_{41}, & Y_{42}, & \ldots, & Y_{4n} & \rightarrow & T_4 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}
$$

Does the practice match the theory?

In theory, if our data is normal, the sample mean is normal. Let's test this.

1. Consider samples of size 3, $Y_1, Y_2, Y_3 \sim N(5, 2^2)$.
2. Every time we take a sample, calculate the mean

$$\bar{Y} = \frac{1}{3}\left(Y_1 + Y_2 + Y_3\right) .$$

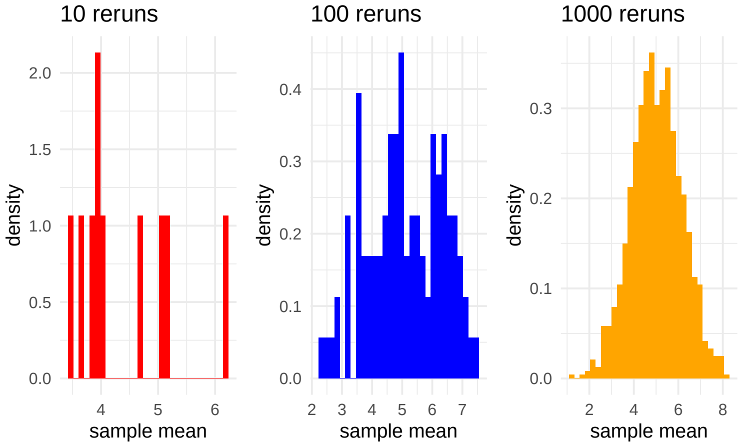
3. Generate 10, 100, and 1000 samples to look at the distribution.
4. Is it normal?

Some R code to do this

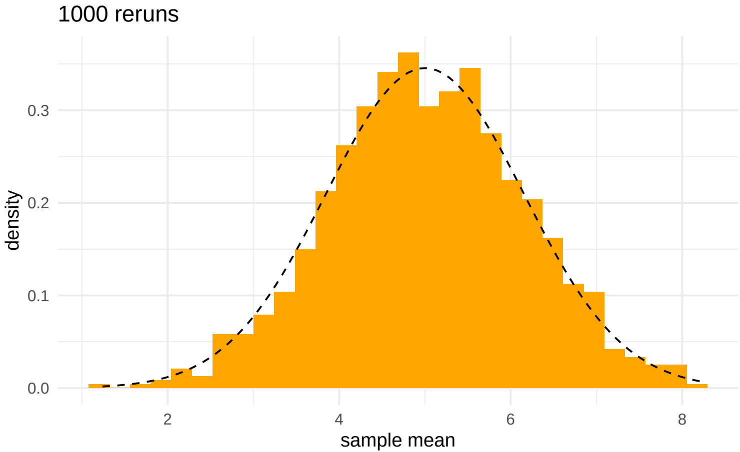
```
# Set up some parameters
N <- 10
mu <- 5
sig <- 2
n <- 3

# Get the samples and calculate the mean
norm_sample_3_10 <- N %>%
  rerun(rnorm(n, mu, sig)) %>%
  map_dbl(mean) #Hey look, a new function!
```

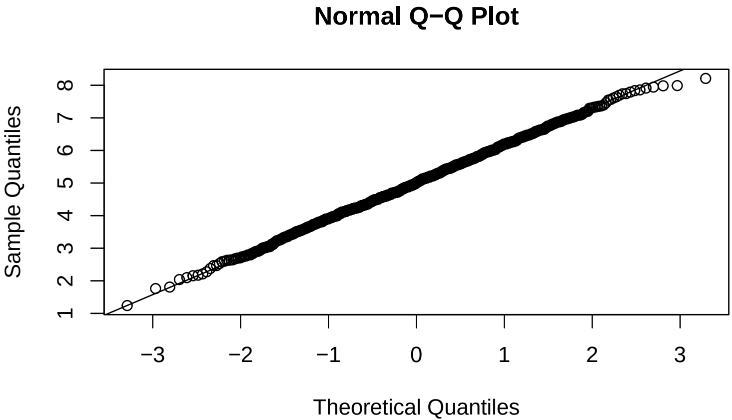
Histograms


10 reruns
100 reruns
1000 reruns
density
2.0
1.5
1.0
0.5
0.0
4
5
6
sample mean
0.4
0.3
0.2
0.1
0.0
2
3
4
5
6
7
0.3
0.2
0.1
0.0
2
4
6
8

Is this normal?



QQplot


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1
2
3
4
5
6
7
8
−3
−2
−1
0
1
2
3

Your turn

What to do

1. Adapt the given code to produce the histograms for $N = 10, 100, 1000$.
2. Explore the distribution as you increase n.
3. Explore the distribution as you change μ and σ^2.

Non-normal data

The problem

Our distributional result relies on the fact that $Y_i \sim N(\mu, \sigma^2)$, although we know

$$\mathrm{E}[\bar{Y}] = \mu$$

and

$$\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n}\,.$$

CLT to the rescue?

Let $Y_1, Y_2, \ldots, Y_n$ be independent independent and identically distributed random variables with $\mathrm{E}[Y_i] = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\,.$$

Then the distribution of U_n converges to the standard normal distribution function as $n \to \infty$.

The problem

The CLT only kicks in for large n, the worse the distribution, the larger the n needed.

χ^2_5

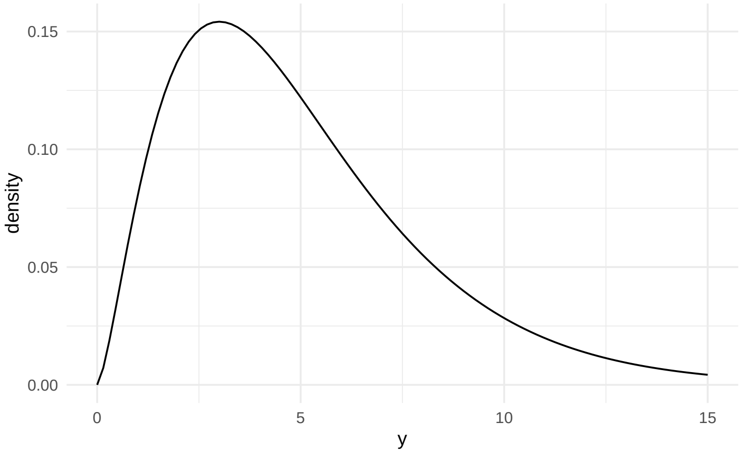
Let's explore the sampling distribution of the sample mean for $Y_1, Y_2, \ldots, Y_n \sim \chi^2_5$. We will

1. Consider samples of size 3, $Y_1, Y_2, Y_3 \sim \chi^2_5$.
2. Every time we take a sample, calculate the mean

$$\bar{Y} = \frac{1}{3}\left(Y_1 + Y_2 + Y_3\right) .$$

3. Generate 10, 100, and 1000 samples to look at the distribution.
4. Is it normal? Expect to see $N(5, 10/3)$.

Is the χ^2_5 normal?

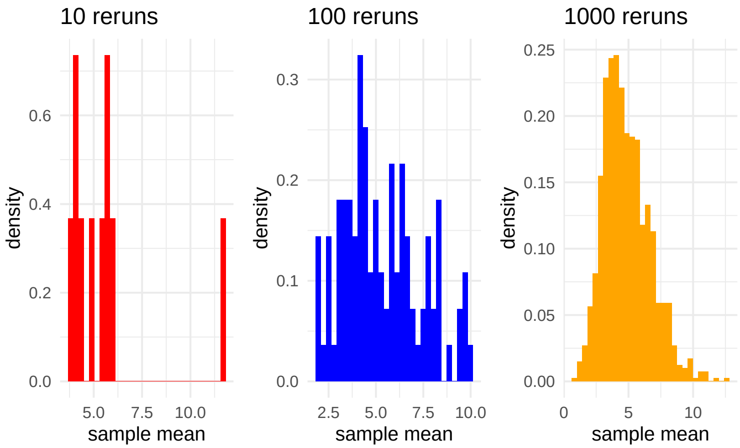


Some R code to do this

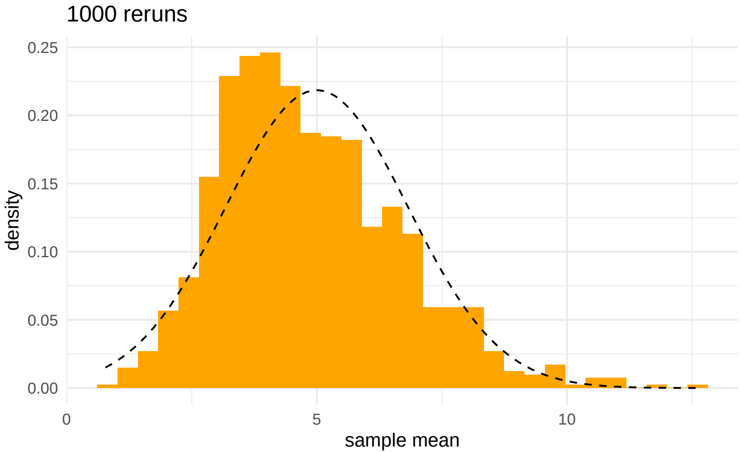
```r
# Set up some parameters
N <- 10
df <- 5
n <- 3

# Get the samples and calculate the mean
chi_sample_3_10 <- N %>%
  rerun(rchisq(n, df)) %>%
  map_dbl(mean)
```

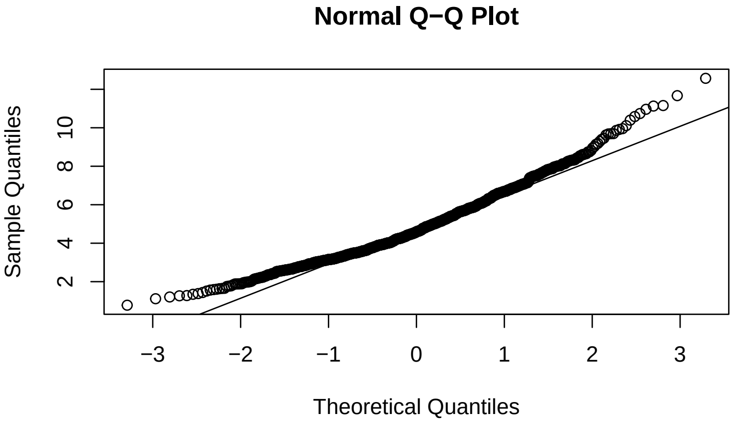
Histograms


10 reruns
100 reruns
1000 reruns
density
0.0
0.2
0.4
0.6
5.0
7.5
10.0
sample mean
0.0
0.1
0.2
0.3
2.5
5.0
7.5
10.0
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10

Is this normal?



QQplot


Normal Q−Q Plot
Sample Quantiles
2
4
6
8
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Your turn

What to do

1. Explore the distribution of the sample mean as you increase the sample size n from the χ^2_5. When does it start to become normal?
2. Look at the t_5 distribution. Explore the sampling distribution of the sample mean. When does it start to become normal?
3. If you had a dataset with no knowledge of its distribution, how might you explore the sampling distribution of the sample mean?