# STATS 2107
# Statistical Modelling and Inference II

## Workshop 5: Sampling distributions part 2

Matt Ryan

School of Mathematical Sciences, University of Adelaide

Semester 2 2022

# The sampling distribution of the P-value

# The Normal hypothesis test

Consider the hypothesis test on $X_1, X_2 \ldots, X_n$ where $X_i \sim N(\mu, \sigma^2)$ and $\sigma^2$ is known. The simple null hypothesis is

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu \neq \mu_0$$

with test statistic

$$Z^* = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

# Let's think about the p-value

By definition, the P-value is

$$P = P(|Z| > z^*)\,,$$

where $z^*$ is the observed value of the test statistic.

***What if I told you this a random variable?***

# The p-value as a random variable

If the $X_i$, $i = 1, 2, \ldots, n$ are not yet observed, then

$$Z^* = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

is random. Hence

$$P = P(|Z| > Z^*)$$

is random

# A thrilling question

*If $P$ is random, what is its distribution?*

# How can we simulate a p-value

To explore the distribution of the p-value, we need 3 things:

1. A null distribution (known values of $\mu$ and $\sigma^2$).
2. Some data from the null distribution.
3. To calculate the P-value.

# A null distribution

Let's suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(0,1)$ for simplicity.
Then the null hypothesis we are testing is

$$H_0 : \mu = 0 \,.$$

# How do we get data?

The easiest way to get data is to simulate it using R. Let's simulate a sample of $n = 100$ observation, which we can do with

```
rnorm(n = 100, mean = 0, sd = 1)
```

# Get the P-value

To do this, we need to calculate the test statistic $z^*$, and calculate

$$P = P(|Z| > z^*) = 2P(Z < -|z^*|)\,.$$

# R code for the p-value

```
x <- rnorm(n = 100, mean = 0, sd = 1)
z <- mean(x)/(1/sqrt(100))
p <- 2*pnorm(-abs(z))
```

# How does this help?

These are the steps to simulate a single p-value. If we do this LOTS and LOTS of times, we can then plot the simulated distribution to see how it looks (with a histogram).

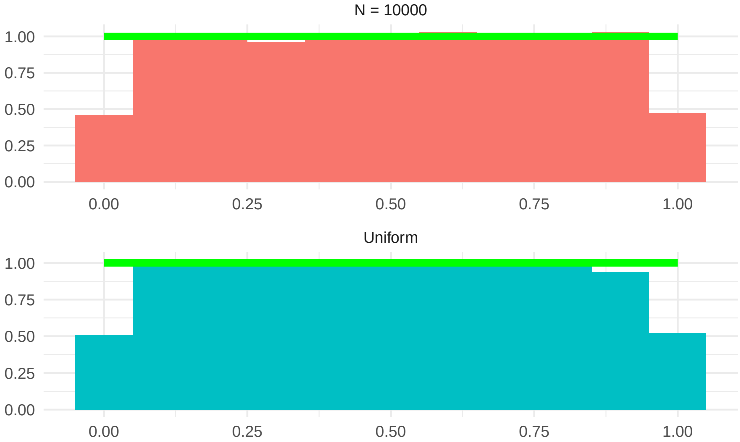
***This is where you come in***

# Your turn

# What to do

1. Write R code to simulate $N = 10$ p-values. Generate a histogram of these p-values. **Hint: Use a for loop**
2. Adapt your code to simulate $N = 100, 1000, 10000$ p-values. Generate histograms for each value of $N$.
3. Propose a sampling distribution for $P$.

# What did I get:



# The sampling distribution of P is uniform

# A powerful theorem

Let $X$ be a continuous random variable with invertible CDF $F(x)$.
Then the random variable $Y = F(X)$ is a $U(0,1)$ random variable.

# Why is this useful

1. The CDF of a continuous random variable is strictly monotonic, hence invertible. Thus this applies to many random variables.
2. This allows us to simulate random variables.

# A proof.

Observe that

$$\begin{aligned} P(Y \leq y) &= P(F(X) \leq y) \\ &= P(X \leq F^{-1}(y)) \\ &= F(F^{-1}(y)) \\ &= y\,. \end{aligned}$$

# A proof

1. CDFs uniquely identify distributions
2. The CDF of $U \sim U(0,1)$ is $F_U(u) = u$.

# How does this help us.

Consider the definition of the P-value:

$$P = P(|Z| > Z^*) = 1 - P(|Z| < Z^*)\,.$$

Then $|Z|$ is a random variable, so

$$1 - P = F_{|Z|}(Z^*)\,.$$

# How does this help us.

Thus $1 - P \sim U(0,1)$, so $P \sim U(0,1)$.

Your turn

# What to do

1. Under the null hypothesis, what is the probability that $P \leq \alpha$? How does this relate to the interpretation of the P-value?
2. How would you use the theorem that if $U = F_Y(y)$, then $U \sim U(0,1)$, to generate random simulations from the distribution $Y$.
3. How does the distribution of the P-value change if the null hypothesis is false?