

Problem Sheet 4

MATH50011
Statistical Modelling 1

Week 5

Lecture 9 (Hypothesis Testing)

1. A clinical trial is conducted to establish whether treatment A has a different effect on systolic blood pressure than treatment B. The treatment effect will be measured by comparing the mean of a group receiving A and a group receiving B. Write a few sentences suitable for a non-statistician (better yet, a non-mathematician!) explaining what is meant by the terms *type I error* and *type II error* in this context.

2. In the results of a study, you read the following sentence:

“Since $p < 0.05$, there is a less than 5% chance that the null hypothesis is true.”

Write an explanation of why this statement is false. Give a correct interpretation of a p-value suitable for a non-statistician.

3. The mean level of prothrombin in the general population is known to be 22.0 mg/100 ml of plasma. A sample of 30 patients showing vitamin K deficiency has a mean prothrombin level of 19.5 mg/100ml and standard deviation 4 mg/100ml of plasma.
 - (a) Is the mean prothrombin level in patients with vitamin K deficiency different from that in the general population? Set up a null and alternative hypothesis that addresses this question.
 - (b) Find the critical values for the sample mean to test the hypotheses in (a) using $\alpha = .05$. Use these critical values to test the hypotheses in (a); state your statistical and scientific conclusions. Clearly state the test statistic and distribution you are using.
 - (c) Compute the p-value and use that to test the hypotheses in (a) using $\alpha = .05$. State both your statistical and scientific conclusions.
4. Suppose that, in the general population, birth weights are approximately normally distributed with a mean weight of 3200g and a standard deviation of 400g. A sample of 25 babies born to teenage mothers has an average birth weight of 2980g.

- (a) You would like to use this sample to determine if the average birth weight of babies born to teenage mothers is different from the general population. Set up a null and alternative hypothesis that addresses this question and carry out a hypothesis test using $\alpha = .05$. Be sure to state both your statistical and scientific conclusions.
- (b) Give a 95% confidence interval for the mean weight of babies born to teenage mothers.
- (c) Describe how you could use the confidence interval from (b) to test the hypotheses in (a).
5. In this exercise, we consider a typical *sample size calculation*. These types of methods are common during the planning of a study.
- Let X_1, \dots, X_n be a random sample from a $N(\theta, \sigma^2)$ population with σ^2 known. We consider testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. We will reject H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > c$.
- (a) Find the power function $\beta(\theta)$ of the test. Express your answer in terms of the standard normal cdf $\Phi(z)$;
- (b) Define the value c_α of c such that the test has level α ;
- (c) Find values c and n such that the test has level α and $\beta(\theta) \geq b$ for all $\theta \geq \theta_0 + \sigma$.
6. **(Challenge)** Binomial data gathered from more than one population are often presented in a *contingency table*. For the case of two populations, the table might look like this:

	Population		
	1	2	Total
Successes	S_1	S_2	$S = S_1 + S_2$
Failures	F_1	F_2	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

where Population 1 is Binomial(n_1, p_1), with S_1 successes and F_1 failures, Population 2 is Binomial(n_2, p_2), with S_2 successes and F_2 failures, and S_1 and S_2 are independent.

We consider testing the hypothesis that $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$.

- (a) Consider the statistic

$$W = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \hat{p}(1 - \hat{p})}$$

where $\hat{p}_k = S_k/n_k$ for $k = 1, 2$ and $\hat{p} = S/n$.

Show that $W \rightarrow_d \chi_1^2$ as $n_1, n_2 \rightarrow \infty$. Explain how an approximate level α test can be constructed based on W .

- (b) We may alternatively measure the departure from H_0 in terms of the difference between the observed frequencies S_1, S_2, F_1, F_2 and the *expected frequencies*:

	Expected Frequencies		
	1	2	Total
Successes	$n_1 S/n$	$n_2 S/n$	$S = S_1 + S_2$
Failures	$n_1 F/n$	$n_2 F/n$	$F = F_1 + F_2$
Total	n_1	n_2	$n = n_1 + n_2$

Consider the statistic

$$W^* = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

where the sum is taken across all 4 cells of the tables.

Show that $W^* = W$ so that W^* is also asymptotically chi squared.

(This is the most common form of the *chi squared test of independence*.)

7. A famous medical experiment was conducted by Joseph Lister in the late 1800s. In his experiment, Lister tested whether carbolic acid (a disinfectant) could reduce the risk of mortality following surgery. Data based on 75 amputations with and without the use of carbolic acid are presented in the following table:

		Carbolic acid used?	
		Yes	No
Patient lived?	Yes	34	19
	No	6	16

Use these data and the test you derived in the previous exercise to test whether the use of carbolic acid is associated with patient mortality.

You may use the R function `prop.test()` to check your by-hand solution.

Lecture 10 (Likelihood Ratio Tests)

8. **(Challenge)** Suppose that X_1, \dots, X_m and Y_1, \dots, Y_n are random samples and X_i is independent of Y_j for all i and j . Suppose further that $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_j \sim N(\mu_y, \sigma_y^2)$ for all i and j . We consider testing $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ for $\theta = \mu_x - \mu_y$.

- (a) Assume that $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Derive the LRT for testing the hypotheses given above. Show that the LRT can be based on the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_P^2 \left(\frac{1}{m} + \frac{1}{n} \right)}}$$

where

$$S_P^2 = \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2 \right).$$

The quantity S_P^2 is called the *pooled variance estimator*.

- (b) Show that, under H_0 , $T \sim t_{n+m-2}$. This yields the *two-sample t-test*.
- (c) Samples of wood were obtained from the core and periphery of a certain Byzantine church. The date of the wood was determined, giving the following data.

```
core      <- c(1294, 1279, 1274, 1264, 1263, 1254, 1251,
1251, 1248, 1240, 1232, 1220, 1218, 1210)
periphery <- c(1284, 1272, 1256, 1254, 1242,
1274, 1264, 1256, 1250)
```

Use the two-sample t-test to determine if the mean age of the core is the same as the mean age of the periphery.

You should complete this exercise by computing the relevant quantities (you may use a calculator). The command `ttest(x,y)` function in R can be used to check your answer.

9. Let Y_1, \dots, Y_n be i.i.d. $\text{Geometric}(p)$ for unknown $p \in (0, 1)$. Then the pmf for each Y_i is

$$P(Y_i = k) = p(1 - p)^{k-1}, \quad k = 1, 2, \dots$$

- Find the maximum likelihood estimator for p .
 - Construct the likelihood ratio test statistic t for testing $H_0 : p = 0.5$ against $H_0 : p \neq 0.5$.
 - Under H_0 , state the asymptotic distribution of t as $n \rightarrow \infty$. You do not need to verify regularity conditions.
 - Describe how to use t to construct an asymptotic level α test.
10. Let X_1, \dots, X_n be a random sample from a parametric model with marginal density function $f_\theta(x)$. Assume below that regularity conditions hold.
- Let $\hat{\theta}_n$ be the MLE of θ_0 . Show that if $\theta = \theta_0$, the *Wald statistic* defined as $W = (\hat{\theta}_n - \theta_0)^2 / \text{SE}(\hat{\theta}_n)^2$ is asymptotically χ_1^2 .
Explain how this can be used to construct an approximate level α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.
 - How would you expect the results of the *Wald test* in part (a) to compare to the results of a likelihood ratio test for large sample sizes? Explain.
 - Sketch a graph of a log-likelihood function and label the MLE $\hat{\theta}_n$ and θ_0 on the horizontal axis. Indicate on your graph the quantities used to calculate the Wald test statistic and likelihood ratio test statistic.

R lab: The Bootstrap

This exercise introduces concepts through use of the R software package.

Let T_n be an asymptotically normal estimator of θ based on a random sample Y_1, \dots, Y_n . We now consider a flexible method called the bootstrap that allows us to approximate the sampling distribution of T_n using observations y_1, \dots, y_n .

The bootstrap sampling distribution can be used to construct confidence intervals for θ by either:

- i. Computing $SE(T_n)$ with respect to the bootstrap sampling distribution and using the formula $T_n \pm c_{\alpha/2} SE(T_n)$ from the notes;
- ii. Computing the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap sampling distribution.

This procedure is widely applicable, but is most useful for estimators where it is difficult to obtain a closed-form expression for $SE(T_n)$.

In R, the code below shows how we usually compute \bar{y} and estimate its standard error based on 4 data points: 2, 4, 9, and 12.

```
y <- c(2,4,9,12)
ybar <- mean(y)
se.ybar <- sqrt(var(y)/4)
```

Running the above code, we find that the standard error is about 2.29.

The bootstrap sampling distribution of \bar{Y} is obtained by resampling the data points with replacement and computing \bar{Y} based on the resampled data. There are 4 data points, so there are $4^4 = 256$ equally likely resamples. We can use R to obtain all 256 values in the bootstrap sampling distribution as follows:

```
# All  $4^4 = 256$  possible resamples with replacement
y.star <- expand.grid(y,y,y,y)

# All 256 sample means based on resampling w/replacement
ybar.star <- apply(y.star, 1, mean)

# The standard error based on this is
se.ybar.star <- sqrt(var(ybar.star))
```

From the above, we find that the bootstrap standard error is about 1.98.

It is a good idea to also visualise the bootstrap distribution of \bar{Y} . This can be achieved with `hist(ybar.star)`. For large sample sizes, we would expect the bootstrap sampling distribution to look approximately normal. The normal approximation for $n = 4$ seems to be less than ideal.

The number of bootstrap resamples n^n grows too quickly to be reasonable for the average statistician. Instead, we usually approximate the bootstrap sampling distribution by drawing a large number of random samples as follows.

```
set.seed(50011)

ybar.boot <- numeric(length = 10000)
for(i in 1:10000){
  y.boot <- sample(y, size = 4, replace = TRUE)
  ybar.boot[i] <- mean(y.boot)
}
se.ybar.boot <- sqrt(var(ybar.boot))
```

From the above, we find that the bootstrap standard error is about 1.99. This is nearly the same as the value obtained by enumerating all 256 samples.

11. Using the code examples above:

- (a) Construct three approximate 95% confidence intervals for the mean μ based on the formula $T_n \pm c_{\alpha/2} \text{SE}(T_n)$ where the standard error is based on `se.ybar`, `se.ybar.star`, and `se.ybar.boot`.
- (b) Construct two additional approximate 95% confidence intervals for the mean with limits defined by the 2.5% and 97.5% percentiles of `ybar.star` and `ybar.boot`. (Hint: use the `quantile()` function.)
- (c) Compare the similarities/differences in the confidence intervals you constructed in parts (a) and (b).
- (d) Replace the data `y` in your code with a random sample of $n = 30$ standard exponential random variables: `y <- rexp(n=30)`. Based on your previous code, construct an approximate 95% confidence interval for the mean based on a normal approximation. Construct two different bootstrap 95% confidence intervals for the mean based on 10000 resamples. (Note: you are not being asked to enumerate all 30^{30} resamples.)