Hypothesis Testing

Introduction and motivation

- So far we have focused on the problem of estimation the unknown population parameter based on the observable data. We wanted to obtain the best possible guess $\hat{\theta} = T(X_1, \dots, X_n)$ of the parameter θ of the distribution in population. In this case we called this problem a **point estimation**, i.e. we obtained a specific point $\hat{\theta} \in \Theta$ in the parametric space.
- Or we wanted to estimate the confidence interval $(L(X), U(X)) \subset \Theta$, such that it contains true parameter value with probability $(1 - \alpha)$, $0 < \alpha < 1$.
- Now we turn to, in some sense, simpler problem: to determine whether the unknown θ belongs to one subset $\Theta_1 \subset \Theta$ or to another $\Theta_2 \subset \Theta$.

Initial setup

- Let $X_1, \dots, X_n \tilde{f}(x|\theta)$, where θ is unknown parameter of the distribution in population. Assume that θ lies in the parameter space Θ, i.e. $\theta \in \Theta$.
- The hypothesis testing problem is specified by splitting the parameter space Θ into two disjoint subsets: $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 =$.
- Then two hypotheses are stated in the following manner:

  $$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_1,$$

  where we call H_0 the *null hypothesis*, and H_1 the *alternative hypothesis*.
- **Goal**: to use only the given sample $X = X_1, \dots, X_n$ to decide between H_0 and H_1.
- We need to construct specific test function $T(X)$, which tells us whether to reject, or do not reject H_0.

  Important idea on this step is that conclusions: a) never say anything directly about H_1, and b) never say that one of two hypotheses is true. We can not conclude just from the sample data that either H_0 or H_1 is true. We may say 'accept H_0', but in fact it means 'do not reject H_0'. Practically there is no difference, but logically the difference is huge: it is impossible for an experiment to confirm a theory, but when enough evidence exists to suggests that a theory is false, it is standard to reject that theory and develop a new one. That's why sometimes H_0 is called a *conservative* hypothesis, which is not to be rejected unless the evidence is clear. On the other hand an alternative H_1 is sometimes referred to as the *research worker's* hypothesis, so it's like an competitive idea, the set of evidences which we would like to test. Test functions $T(X)$ are like estimators in the estimation problem, and our primary focus is how to choose a "good" test.

There are four possible outcomes in hypothesis testing problem:

	H_0 is true	H_0 is false
Accept H_0	correct	Type II error
Reject H_0	Type I error	correct

For instance we could decide that $\theta \in \Theta_1$ when really $\theta \in \Theta_0$, that is Type I error, or we could decide that $\theta \in \Theta_0$ when, in fact, $\theta \in \Theta_1$, which is Type II error respectively. Still remember that "accept H_0" actually means "Do not reject H_0 ."

Basics

The two primary characteristics of a test $T(X)$ are size and power.

- Let us for simplicity assume that $T(X)$ is binary test, which returns 1, if this is 'to reject H_0' and 0 if 'do nor reject H_0'.
- Don't forget that $T(X)$ is statistic, function from the given sample. It means that on one sample it may return 0, but on the one 'bad' sample it may return 1.

- Let us denote by C set of **all samples of equal length** such that $T(X) = 1$, $\forall X \in C$. The set C is called *critical region.* The specific test $T(X)$ determines its own critical region.
- Test $T(X)$ also determines the anti-friend of C, which we denote as $\bar{C}$, namely the complement. The set $\bar{C}$ is called the *acceptance* region, the set of all possible samples where test $T(X)$ says 'do not reject H_0.'
- **Definition**: Given Θ_0, the *size* of the test $T(X)$ is:

$$\text{size} = \max_{\theta \in \Theta_0} P\{X \in C|\theta\}.$$

In other words size of the test is the probability of Type I error. You can argue that the probability of an error has nothing common with the sample,and this is a characteristic of specific test function $T(X)$, so the test itself is bad. But recall that result of a test is uniquely determined by a sample, because test is a function from the sample. This mean that for a given test function $T(X)$ the probability of Type I error is the same as the probability to get a 'bad' sample $X \in C$.

- Size of the test $T(X)$ is also often called the *significance level* of the test and denoted by α.
- **Definition**: Given the test function $T(X)$, the *power function* of the test is:

$$W(\theta) = P\{X \in C|\theta\}.$$

Here can be noticed that $s\alpha = \max_{\theta \in \Theta_0} W(\theta)$, and in the case $\theta \in \Theta_1$ we have:

$$1 - W(\theta) = P\{X \in \bar{C} | \theta\} = P(\text{type II error}).$$

- Clearly, we want *significance level* to be small. For *power* we would like to see more how it behaves for θ outside of Θ_0: there $W(\theta)$ denotes the probability of correctly rejecting H_0 when it's false, in this light we want *power* to be large

Finally, the goal statement:

1. We want to fix the size at some small value, usually $\alpha = 0.05$, and then seek for a test that has this size and maximizes the power $W(\theta)$.
2. Why we fix α and maximize $W(\theta)$ is that two quantities are competing: if we force the test to have small size, then we indirectly reduce the power, and vice versa. That is, improving in one area hurts in the other. So the accepted strategy is to allow a little chance of a Type I error with the hopes of making the power large.

Hypotheses about parameters of the distribution:

Problem 1

Let $X_1, \ldots, X_n \sim N(\theta, 1)$, and consider $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$.

(a) Define a test with critical region $C = (X_1, \ldots, X_n) : \bar{X} > k$. Find k such that the size of the test is $\alpha = 0.05$.

(b) For the test you derived in (a), find the power function $W(\theta)$.

Problem 2

A particular car is advertised to have gas mileage 30mpg. Suppose that the population of gas mileages for all cars of this make and model looks like a normal distribution $N(\mu, \sigma^2)$, but with both μ and σ unknown. To test the advertised claim, sample $n = 10$ cars at random and measure their gas mileages $X_1, \ldots, X_{10}$. Use these data to test the claim, i.e., test $H_0 : \mu = 30$ vs. $H_1 : \mu < 30$. Assume $\bar{X} = 26.4$ and $S = 3.5$ are the observed sample mean and standard deviation, respectively. Use $\alpha = 0.05$.

Continuation: other situations and distributions

Sample proportions

- Let us briefly recall how we get error margins for CI, and how we use them in HT

- Remember the density function of the normal distribution:
$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \sim \mathcal{N}(\mu, \sigma^2).$$

- Also recall the formula for probability in binomial distribution:
$$P\{X = x\} = C_n^x p^x q^{n-x} \sim \text{Bin}(n, p).$$

- Important **de Moivre–Laplace theorem** states that normal distribution may be used as an approximation to the binomial distribution under certain conditions. Namely: as n grows large, for x in the neighborhood of np we can approximate:
$$C_n^x p^x q^{n-x} \approx \frac{1}{\sqrt{2\pi np(1-p)}}e^{-\frac{(x-np)^2}{2np(1-p)}} \sim \mathcal{N}(np, np(1-p)).$$

We use that property when constructing confidence intervals and performing hypotheses testing on population proportion.

- From the property above we can obtain that within the sample size large enough:
$$\hat{p} \sim \mathcal{N}(p, \frac{p(1-p)}{n}), \quad n > 30.$$

- Let us then transform random variable to standard one as we used to do::
$$Z = \frac{X - \mu}{\sigma} \longrightarrow \frac{\hat{p} - p}{\sqrt{pq/n}} \sim \mathcal{N}(0, 1).$$

- From that point it is clear how we got error margin for CI. Imagine we found the $z_{\alpha/2}$, as an upper boundary. Then:
$$\frac{\hat{p} - p}{\sqrt{pq/n}} < z_{\alpha/2} \rightarrow \hat{p} - z_{\alpha/2}\sqrt{\frac{pq}{n}} < p \rightarrow\approx z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p,$$

which gives us well known $\hat{p} - E_{\hat{p}}$, where we substitute p for $\hat{p}$, because it's the only estimation that we have.

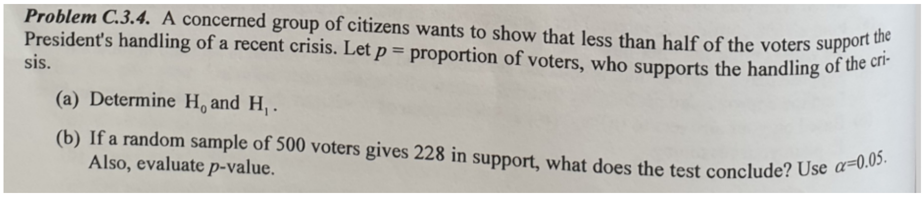
Proportions in Hypothesis Testing:

When we do hypothesis testing $H_0 : p = p_0$ versus $H_1 : p < (>)p_0$, we can substitute $\hat{p}$ (sample proportion) by p_0 (population proportion that we want to test), or stay with $\hat{p}$:

$$Z_1 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < (>)c, \text{ or } \quad Z_2 = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} < (>)c.$$

When n is large both Z_1 and Z_2 have approximate standard normal distributions provided that $H_0 : p = p_0$ is true.

Problem 3



Problem C.3.4. A concerned group of citizens wants to show that less than half of the voters support the President's handling of a recent crisis. Let p = proportion of voters, who supports the handling of the crisis.

(a) Determine H_0 and H_1.

(b) If a random sample of 500 voters gives 228 in support, what does the test conclude? Use α=0.05. Also, evaluate p-value.

Mean differences

Sometimes we need to compare two samples between each other. - Were they obtained from the same distribution? - Was there any difference after changing some parameter?

In the course namely we would be intrested in comparing populations means μ_1 and μ_2 of the sampling distributions of means $\bar{X}$ and $\bar{Y}$.

There are a few different situations, which have different approaches to tackle with them.

Test on means equality, known variances

Assume we have two independent samples: $X = X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$, and $Y = Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

- We are interested in the behaviour of the parameter $\theta = \mu_1 - \mu_2$.
- Let's introduce $\hat{\theta} = \bar{X} - \bar{Y}$, the estimator of θ.
- $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_1 - \mu_2 = \theta$, so $\hat{\theta}$ is an unbiased estimator of θ.
- As X and Y are independent we can write formula connecting their variances:

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

- Remember the CLT? :

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

That also would work for the new variable $\hat{\theta}$:

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \sim N(0,1).$$

- By this, you can construct confidence intervals or perform hypothesis testing on the equality of population means, namely; $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ (two-tailed test) or $H_1 : \mu_1 > (<)\mu_2$ (one-tailed test).

- Sometimes things can be even more simple if it is given that $\sigma_1 = \sigma_2 = \sigma$ and/or $m = n = l$. That will simplify the formula a lot!

Test on means equality, unknown equal variances

As we discussed, in many cases true values of σ_1 and σ_2 may not be known to us. We are again coming to the **Student's t-distribution**!

Consider slightly changed previous setting: we have two independent samples with equal variances, but they are unknown:

$X = X_1, \dots, X_n \sim \mathcal{N}(\mu_1, \sigma^2)$, and $Y = Y_1, \dots, Y_m \sim \mathcal{N}(\mu_2, \sigma^2)$.

- Recall how we construct a t-distributed variable?

$t = \frac{Z}{\sqrt{\frac{\chi^2(k)}{k}}}$, where Z is a standard normal variable, and k is a number of degrees of freedom.

- So on one hand we have:

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} = \frac{\hat{\theta} - \theta}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1).$$

- On the other hand we can construct a new χ^2 variable::

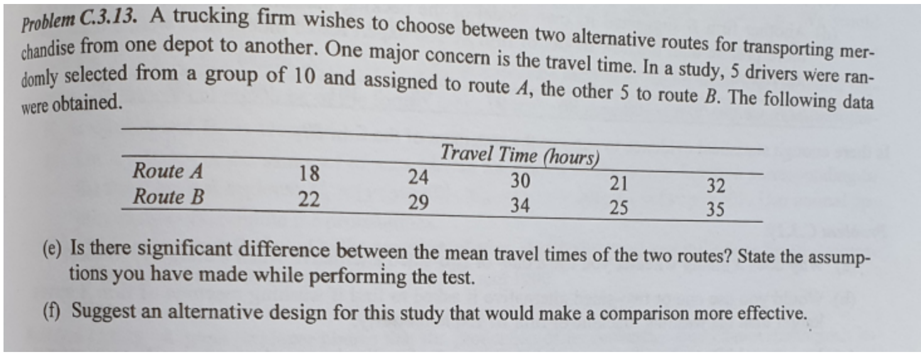
$$\frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n + m - 2).$$

- Finally, construct a t-distributed variable!

$T = \frac{\hat{\theta} - \theta}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}\frac{S}{\sigma}} = \frac{\hat{\theta} - \theta}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n{+}m{-}2)$, where S is a **pooled variance:** $S = \frac{(n-1)S_x^2 + (m-1)S_Y^2}{n + m - 2}$

- By this, again, you can construct confidence intervals or perform hypothesis testing on the equality of population means, namely; $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ (two-tailed test) or $H_1 : \mu_1 > (<)\mu_2$ (one-tailed test).
- Things can also be simplified if $m = n = l$.

Problem 4



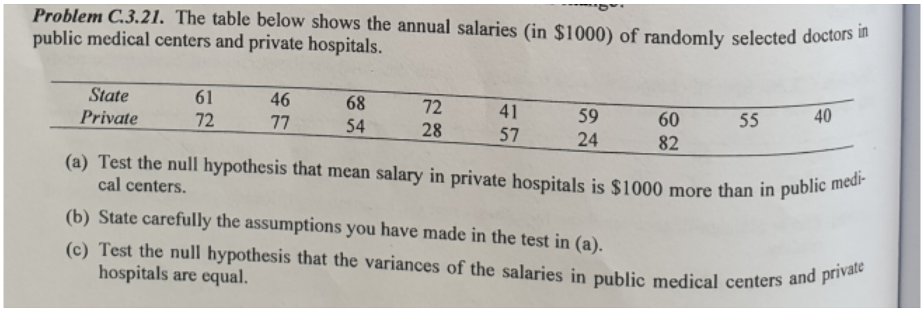
Problem C.3.13. A trucking firm wishes to choose between two alternative routes for transporting merchandise from one depot to another. One major concern is the travel time. In a study, 5 drivers were randomly selected from a group of 10 and assigned to route A, the other 5 to route B. The following data were obtained.

	Travel Time (hours)				
Route A	18	24	30	21	32
Route B	22	29	34	25	35

(e) Is there significant difference between the mean travel times of the two routes? State the assumptions you have made while performing the test.

(f) Suggest an alternative design for this study that would make a comparison more effective.

Problem 5



Problem C.3.21. The table below shows the annual salaries (in \$1000) of randomly selected doctors in public medical centers and private hospitals.

State	61	46	68	72	41	59	60	55	40
Private	72	77	54	28	57	24	82		

(a) Test the null hypothesis that mean salary in private hospitals is \$1000 more than in public medical centers.

(b) State carefully the assumptions you have made in the test in (a).

(c) Test the null hypothesis that the variances of the salaries in public medical centers and private hospitals are equal.