Uppsala Universitet
Matematiska Institutionen
I. Kaj, R. Müller, Y. Wang

Sannolikhet och statistik DV, 1MS321
Lektion 8

Repetition och viktiga formler för M8

- Understand the definition of confidence interval
- Calculate CI of expectations for normal and binomial distribution
- Understand t-distribution
- Calculate CI of difference between expectations

1 Definition and Interpretation of CI

"Why do we need to study confidence intervals?"

In chapter 6, we have seen point estimation. Basically what we did with a set of sampled data is, assuming it was generated from a probability distribution, and estimated the unknown parameters of the distribution. However, the estimation problem is not yet completely solved, because there is still uncertainty in our estimation. Consider the following two statements, for example:

1. **Fortune Teller:** The amount of rain tomorrow is 25mm.
2. **Scientist:** I believe that the amount of rain tomorrow is 25mm, plus minus 3mm, and I'm 90% sure of my prediction.

This is essentially the difference between giving the point estimator and giving the point estimator AND a corresponding confidence interval. To behave like a scientist, we need to answer the following questions:

- How good is our estimate?
- How can we measure its quality?

A *confidence interval* (CI) is an interval used to estimate the likely size of a parameter. A *confidence level* is a measure of the degree of reliability of the confidence interval.
Let A, B be functions of $X_1, X_2, \dots, X_n$, such that

$$\mathbb{P}(A \leq \theta \leq B) = 1 - \alpha,$$

then $[A, B]$ is called $100(1-\alpha)$ **percent confidence interval** for θ, with **confidence level** $1-\alpha$.
Most commonly used confidence levels are the 90%, 95% and 99% confidence intervals that have 0.90, 0.95 and 0.99 (corresponding to $\alpha = 0.1, 0.05, 0.01$) probabilities respectively of containing the parameter. For population parameter μ, 95% confidence interval $(\hat{\mu}_d, \hat{\mu}_u)$ of μ is an interval that satisfies

$$P(\hat{\mu}_d \leq \mu \leq \hat{\mu}_u) = 0.95.$$

We usually make the interval *centered* so that

$$P(\hat{\mu}_d \leq \mu) = P(\mu \leq \hat{\mu}_u) = 0.025.$$

Here the subscripts d, u correspond to down and up.

2 The Student's t-distribution

The t-distribution (Student's t-distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

"Why Use the t-Distribution?"

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is **sufficiently large**. Then using the standard deviation of the population, we can use the normal distribution to evaluate probabilities with the sample mean. But what if the sample size is small? Or if we do not know the standard deviation?
When either of these problems occur, statisticians rely on the distribution of the t statistic (also known as the t score), whose values are given by:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

where s is an estimation of the unknown σ, which is given by

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}.$$

We say that T follows the **t-distribution** with n-1 degrees of freedom, $T \sim t(n-1)$.
One might notice that there is a different t-distribution for each sample size, in other words, a t distribution is characterized by its degree of freedom. The t density curves are symmetric

and bell-shaped – behaves very much like the normal distribution and have their peak at 0. However, the spread is always more than that of the standard normal distribution. See Figure 1 below, where $\nu = n - 1$ is the degree of freedom of the distribution.

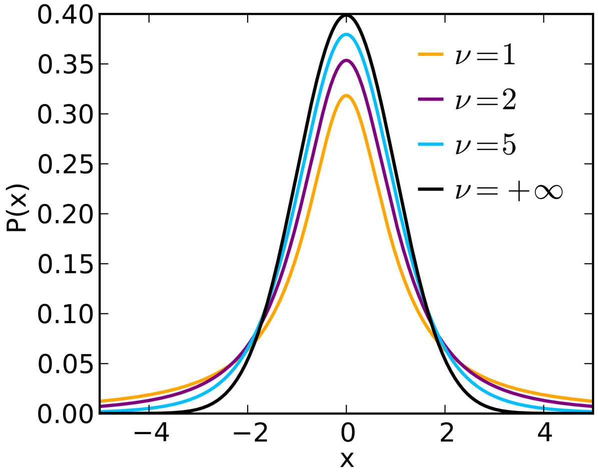

Figure 1: The probability density function of t-distributions (Wikipedia)

Some properties of the t-distribution are:

- t-distribution is different for different sample sizes.
- t-distribution is generally bell-shaped, but with smaller sample sizes shows increased variability (flatter). In other words, the distribution is less peaked than a normal distribution and with thicker tails.
- The mean is zero and the distribution is symmetrical about the mean.
- The variance is greater than one, but approaches 1 as the degree of freedom increases
- As the sample size increases, the distribution approaches a normal distribution. For n $>$ 30, the differences are negligible.

To see a graphical illustration of the t-distribution and a comparison with the standard normal distribution one can see here.

3 CI for the Expectation

First of all, the confidence interval in general can be written in the following form:

$$I = [\hat{\theta} \pm \text{quartile} \cdot d].$$

where $\hat{\theta}$ is your point estimator, which in this case is the sample mean $\bar{x}$. The "quartile" parameter estimates how accurate you want this estimation to be, representing your confidence level, and the d parameter describes how spread out your sample data is. Under different scenarios, we choose different quartile parameters and d parameters.

3.1 Normal distribution

Consider an independent sample $x_1, x_2, \dots, x_n$, coming from $X_1, X_2, \dots, X_n$. We have already calculated its sample mean

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n).$$

Now we want to calculate a $100(1-\alpha)$ percent confidence interval.

- **Scenario 1: Normal sample, σ known**
  In this case take the quartile parameter to be $\lambda_{\frac{\alpha}{2}}$, and $d = \frac{\sigma}{\sqrt{n}}$, so we have

$$I_\mu = [\bar{x} - \lambda_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + \lambda_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}].$$

- **Scenario 2: Normal sample, σ unknown**
  In this case take the quartile parameter to be $t_{\frac{\alpha}{2}}(n-1)$, and $d = \frac{s}{\sqrt{n}}$, so we have

$$I_\mu = [\bar{x} - t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}],$$

  where

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}}.$$

  OBS: The $n-1$ in $t_{\frac{\alpha}{2}}(n-1)$ means its degree of freedom, it does not mean to multiply the parameter by $(n-1)$.

- **Scenario 3: Large sample, not necessarily normal**
  By the central limit theorem, we can assume that the sample comes from a normal distribution, hence we can use the normal distribution quartile parameter: $\lambda_{\frac{\alpha}{2}}$, and we use $d = \frac{s}{\sqrt{n}}$

3.1.1 Length of the CI

In this chapter we assume a symmetric confidence interval, so its length L would be

$$L = 2 \cdot \text{quartile} \cdot d.$$

In the normal case, we would have

$$L = 2\lambda_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

The bigger the sample size, the smaller the length of CI. Therefore, to have the length of the CI no larger than a certain number L, we would require a larger sample size n such that

$$n \geq \left(\frac{2\lambda_{\frac{\alpha}{2}}\sigma}{L} \right)^2.$$

3.2 Binomial Distribution

In the Binomial distribution, the parameter we want to estimate is p. The standard error

$$d = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Therefore the confidence interval can be written as

$$I_p = [\hat{p} - \lambda_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \lambda_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

This method does not work so well with small or large p, hence we have modifications, like the Agresti and Coull interval method, see @JR,

4 Confidence interval for differences in μ

In this section we discuss two main cases: two independent samples or pairwise observations. To tell these two cases apart, the crucial step is to identify how the data was collected, if the sample data we are comparing are independent.

4.1 Two Independent Samples

We have two groups of sample data:

$$x_1, \ldots, x_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2),$$
$$y_1, \ldots, y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

We want to calculate a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$, and see if it contains 0. (If so, we would say that with a certain confidence level, there is a difference between the means.)

- **With known variances**
  In this case we know exactly what the values of σ_1, σ_2 are, and we can write the CI as

$$I_{\mu_1-\mu_2} = \left[\bar{x} - \bar{y} \pm \lambda_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}} \right].$$

- **With unknown variances**
  In this case we assume $\sigma_1 = \sigma_2$. First we will introduce the weighted variance s_p^2:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

  And we write the confidence interval as

$$I_{\mu_1-\mu_2} = \left[\bar{x} - \bar{y} \pm t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

An example of independent samples would be that, if we want to take water samples from lake A and lake B, which are far away from each other.

4.2 Pairwise Observations

In this case we have the data appearing in pairs:

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Let $z_i = x_i - y_i$, for $i = 1, \ldots, n$, and $Z \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_z^2)$. Then we are back in the case where we need to calculate the CI for one set of normally distributed sample data.
An example of pairwise observations would be in medical science, two groups of data representing before and after treatment. In this case they always have the same sample size, and usually have the same object.