**476 Statistics, Spring 2022.**

**Lecturer: Sergey Nadtochiy.**

**Lecture 3. Confidence intervals. (Sections 8.5–8.6, 8.8–8.9)**

**Def 1.** *A two-sided confidence interval for $\theta$ of the confidence level (or, with confidence coefficient) $1-\alpha$ is a pair of statistics $\hat{\theta}_L \leq \hat{\theta}_U$ s.t.*

$$\mathbb{P}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1-\alpha.$$

Analogously we define one-sided confidence intervals.

It is important to remember that $\theta$ is constant, but **a confidence interval is random – it depends on the realizations of the sample!**

A general method to find a confidence interval is called pivotal – it is based on finding a **pivot**.

**Def 2.** *A function of the sample and the unknown parameter $\theta$ is called a* **pivot** *if its distribution does not depend on any unknown parameters (including $\theta$).*

Once we found a pivot $G(Y_1,\dots,Y_n;\theta)$, we aim to find $a(y_1,\dots,y_n) \leq b(y_1,\dots,y_n)$ and a set $S$, s.t. for all $(y_1,\dots,y_n,\theta)$:

1. $a(y_1,\dots,y_n) \leq \theta \leq b(y_1,\dots,y_n)$ if and only if $G(y_1,\dots,y_n;\theta) \in S$,
2. and $\mathbb{P}(G(Y_1,\dots,Y_n;\theta) \in S) = 1-\alpha$.

Typically, $S$ is also an interval, so the item 2 above can be ensured using the quantiles of the distribution of $G(Y_1,\dots,Y_n;\theta)$ (which is fixed, in the sense that it does not depend on any unknown parameters), provided this distributions belongs to one of the standard families (e.g., normal, T, etc.).

Once the desired $S$ and $a, b$ are found, we define the confidence interval as follows:

$$\hat{\theta}_L = a(Y_1,\dots,Y_n), \quad \hat{\theta}_U = b(Y_1,\dots,Y_n).$$

**Ex 1.** *$Y$ is a sample of size 1 from $Exp(1/\theta)$. Find a confidence interval for $\theta$ with confidence level $0.9$.*

$$G(y;\theta) := y/\theta, \quad G(Y;\theta) \sim Exp(1).$$

$$a \leq \theta \leq b \text{ if and only if } y/b \leq G(y;\theta) \leq y/a$$

*Choose $\tilde{a} \leq \tilde{b}$ and $S := [\tilde{a}, \tilde{b}]$, s.t.*

$$\mathbb{P}(\tilde{a} \leq G(Y;\theta) \leq \tilde{b}) = 0.9,$$

*and set*

$$a(y) := y/\tilde{b}, \quad b(y) := y/\tilde{a}.$$

$$0.05 = \mathbb{P}(G(Y;\theta) > \tilde{b}) = e^{-\tilde{b}}, \quad \tilde{b} = -\log 0.05 \approx 2.996,$$

$$0.05 = \mathbb{P}(G(Y;\theta) < \tilde{a}) = 1-e^{-\tilde{a}}, \quad \tilde{a} = -\log 0.95 \approx 0.051,$$

$$\hat{\theta}_L = a(Y_1) = Y_1/\tilde{b} \approx Y_1/2.996, \quad \hat{\theta}_U = b(Y_1) = Y_1/\tilde{a} \approx Y_1/0.051.$$

# 1 Confidence intervals for the mean and variance of a Gaussian sample

In this section, we assume that $Y_1, \dots, Y_n$ is a sample from $N(\mu, \sigma^2)$.

First, let us assume that $\sigma$ is known and let's construct a confidence interval for $\mu$. One way to do it is as follows:

$$G(Y_1; \mu) = (Y_1 - \mu)/\sigma \sim N(0,1),$$

$$a \le \mu \le b \text{ if and only if } (y_1 - b)/\sigma \le G(y_1; \mu) \le (y_1 - a)/\sigma,$$

$$\mathbb{P}(-z_{\alpha/2} \le G(Y_1; \mu) \le z_{\alpha/2}) = 1 - \alpha,$$

$$\hat{\theta}_L = Y_1 - \sigma z_{\alpha/2}, \quad \hat{\theta}_U = Y_1 + \sigma z_{\alpha/2},$$

where $z_{\alpha/2}$ is the quantile of a standard normal at level $1 - \alpha/2$ (NOT a quantile at level $\alpha/2$ – to be consistent with the book). The length of the interval is

$$2\sigma z_{\alpha/2}$$

We can get a smaller confidence interval. Let us consider

$$G(Y_1, \dots, Y_n; \mu) = \sqrt{n}(\bar{Y} - \mu)/\sigma \sim N(0,1),$$

$$a \le \mu \le b \text{ if and only if } \sqrt{n}(\bar{y} - b)/\sigma \le G(y_1, \dots, y_n; \mu) \le \sqrt{n}(\bar{y} - a)/\sigma,$$

$$\mathbb{P}(-z_{\alpha/2} \le G(Y_1, \dots, Y_n; \mu) \le z_{\alpha/2}) = 1 - \alpha,$$

$$\hat{\theta}_L = \bar{Y} - \sigma z_{\alpha/2}/\sqrt{n}, \quad \hat{\theta}_U = \bar{Y} + \sigma z_{\alpha/2}/\sqrt{n}.$$

The length of the interval is

$$2\sigma z_{\alpha/2}/\sqrt{n}$$

Naturally, the length is reduced as we use more information.

Next, assume that $\sigma$ is unknown. Consider

$$G(Y_1, \dots, Y_n; \mu) = \sqrt{n}(\bar{Y} - \mu)/S \sim T(n-1),$$

$$a \le \mu \le b \text{ if and only if } \sqrt{n}(\bar{y} - b)/S \le G(y_1, \dots, y_n; \mu) \le \sqrt{n}(\bar{y} - a)/S,$$

$$\mathbb{P}(-t_{\alpha/2} \le G(Y_1, \dots, Y_n; \mu) \le t_{\alpha/2}) = 1 - \alpha,$$

$$\hat{\theta}_L = \bar{Y} - St_{\alpha/2}/\sqrt{n}, \quad \hat{\theta}_U = \bar{Y} + St_{\alpha/2}/\sqrt{n},$$

where $t_{\alpha/2}$ is the quantile of $T(n-1)$ at level $1 - \alpha/2$. The length of the interval is

$$2St_{\alpha/2}/\sqrt{n}$$

Recall that, because of the heavier tails of the T-distr., $t_{\alpha/2} > z_{\alpha/2}$, which means that the confidence interval constructed without the knowledge of $\sigma^2$ is wider. This is natural: the more information we have, the smaller is the confidence interval. Note, however, that $t_{\alpha/2}$ and $z_{\alpha/2}$ become closer as $n$ increases.

**Ex 2.** *Daily numbers of visits to a website have been recorded for* $8$ *days:*

$$3005,\ 2925,\ 2935,\ 2965,\ 2995,\ 3005,\ 2937,\ 2905$$

**Q 1.** *Assuming that the sample is normal, find a* $95\%$ *confidence interval for the true mean of the daily number of visits.*

*Following the formula established above, we obtain:*

$$\hat{\theta}_L = \bar{y} - st_{\alpha/2}/\sqrt{n} = 2959 - 2.365(39.1/\sqrt{8}) = 2959 - 32.7,$$

$$\hat{\theta}_U = \bar{y} + st_{\alpha/2}/\sqrt{n} = 2959 + 2.365(39.1/\sqrt{8}) = 2959 + 32.7.$$

Next, we ask the following question.

**Q 2.** *What if we have two (indep.) normal samples and we want to construct a confidence interval for $\mu_1 - \mu_2$?*

If $\sigma_i$'s are known, it can be done by using $\bar{Y} - \bar{Z} - (\mu_1 - \mu_2) \sim N(0, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ (recall that a some of independent normals is normal), to obtain

$$\hat{\theta}_L = \bar{Y} - \bar{Z} - z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \quad \hat{\theta}_U = \bar{Y} + z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

If $\sigma_i$'s are not known but coincide, $\sigma_1 = \sigma_2 = \sigma$, then, we can still obtain an exact confidence interval. The main idea is to use the so-called **pooled estimator** for $\sigma^2$:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

**Rem 1.** *It is easy to check that $S_p^2$ is an unbiased and consistent estimator of $\sigma^2$.*

**Rem 2.** *If $n_1 = n_2$, we can construct a confidence interval for $\mu_1 - \mu_2$ by considering the sample $\{U_i := Y_i - Z_i\}$ and noticing that $\mu_1 - \mu_2$ is the mean of $U_i$, while the variance of $U_i$ is $2\sigma^2$. Note also that $2S_p^2$ is* **not** *equal to the sample variance of $\{U_i\}$, but the two become closer as $n \to \infty$.*

Notice that

$$\frac{n_1 + n_2 - 2}{\sigma^2}S_p^2 = \frac{n_1 - 1}{\sigma^2}S_1^2 + \frac{n_2 - 1}{\sigma^2}S_2^2 \sim \chi^2(n_1 - 1) + \chi^2(n_2 - 1) \sim \chi^2(n_1 + n_2 - 2)$$

It also follows from a theorem we had before that $S_p^2$ is independent of $(\bar{Y}, \bar{Z})$. Then,

$$G(Y_1, \ldots, Y_n; \mu_1 - \mu_2) := \frac{\bar{Y} - \bar{Z} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} = \frac{(\bar{Y} - \bar{Z} - (\mu_1 - \mu_2))/\sqrt{\sigma^2/n_1 + \sigma^2/n_2}}{\sqrt{\left(\frac{n_1+n_2-2}{\sigma^2}S_p^2\right)/(n_1 + n_2 - 2)}} \sim T(n_1 + n_2 - 2)$$

This yields the following confidence interval for $\mu_1 - \mu_2$:

$$\hat{\theta}_L = \bar{Y} - \bar{Z} - S_p t_{\alpha/2}\sqrt{1/n_1 + 1/n_2}, \quad \hat{\theta}_U = \bar{Y} - \bar{Z} - S_p t_{\alpha/2}\sqrt{1/n_1 + 1/n_2}.$$

**Ex 3.** *In this example, we consider the testing of a new method of training employees. The performance of an employee is measured in the number of minutes required to assemble a device. Two groups of 9 ppl, one with standard training (denoted $\{Y_i\}$), and one with new training (denoted $\{Z_i\}$), are chosen at random. The realizations of these samples are:*

$$\{y_i\}: 32, 37, 35, 28, 41, 44, 35, 31, 34,$$

$$\{z_i\}: 35, 31, 29, 25, 34, 40, 27, 32, 31.$$

**Q 3.** *Assuming that the sample is normal and that the two samples have the same variance, find a confidence interval for $\mu_1 - \mu_2$ of the confidence level $0.95$.*

*Using the above formula, with the pooled estimator, we obtain:*

$$\bar{y}_1 = 35.22, \quad \bar{y}_2 = 31.56,$$

$$s_1^2 = 24.445, \quad s_2^2 = 20.027,$$

$$s_p^2 = \frac{8 \cdot 24.445 + 8 \cdot 20.027}{9 + 9 - 2} = 22.236, \quad s_p = 4.716,$$

$$\bar{y}_1 - \bar{y}_2 \pm s_p t_{0.025} \sqrt{1/n_1 + 1/n_2} = 35.22 - 31.56 \pm 4.716 \cdot 2.120 \sqrt{1/9 + 1/9} = 3.66 \pm 4.71.$$

*At the confidence level* $0.95$*, we cannot rule out that the new method does not improve the employees' performance.*

Next, we ask the following question.

**Q 4.** *What is the confidence interval for* $\sigma^2$*?*

$$G(Y_1, \ldots, Y_n; \sigma^2) = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1),$$

$$a \le \sigma^2 \le b \quad \text{if and only if} \quad \frac{n-1}{b} s^2 \le \frac{n-1}{\sigma^2} s^2 \le \frac{n-1}{a} s^2$$

Hence,

$$a = \frac{n-1}{\chi^2_{\alpha/2}} s^2, \quad b = \frac{n-1}{\chi^2_{1-\alpha/2}} s^2,$$

where $\chi^2_{\alpha/2}$ is the quantile of $\chi^2(n-1)$ at the level $1 - \alpha/2$.

**Ex 4.** *In this example, we test the variability of a device that measures the volume of an audio source. We apply this device to the same audio source 3 times and obtain the following realized sample values:* $4.1$*,* $5.2$*,* $10.2$*.*

**Q 5.** *Assuming that the sample is normal (i.e., that each measurement made by the device equals the true volume* $\mu$ *plus a normally distributed zero-mean noise, independent across measurements), find a* $0.9$*-confidence interval for the true variance* $\sigma^2$ *of the measurements.*

*Using the above formula, we obtain:*

$$s^2 = 10.57, \ n = 3, \ \alpha = 0.1,$$

$$\chi^2_{0.05} = 5.991, \quad \chi^2_{0.95} = 0.103,$$

$$\left( \frac{3-1}{5.991} 10.57, \ \frac{3-1}{0.103} 10.57 \right) = (3.53, \ 205.25)$$

*The confidence interval is huge because the sample is small and the last measurement is way off.*

## 2 Asymptotic (large-sample) confidence intervals

In this section, we do NOT assume normality of the sample. Then, typically, the results become asymptotic, in the sense that the probability that the true parameter belongs to the confidence interval only approaches $1-\alpha$ as $n \to \infty$.

**Q 6.** *Given an unbiased and asymptotically normal estimator $\hat{\theta}$ for $\theta$, how to construct an asymptotic $(1-\alpha)$-confidence interval for $\theta$?*

Assume that $\sqrt{n}(\hat{\theta}-\theta) \to N(0,\tilde{\sigma}^2)$ as $n \to \infty$. Then,

$$G(Y_1,\dots,Y_n;\theta) := \sqrt{n}(\hat{\theta}-\theta)/\tilde{\sigma}$$

is approximately $N(0,1)$. Using $G$ as the pivot, we follow the standard algorithm:

$$a \le \theta \le b \quad \text{if and only if} \quad \sqrt{n}(\hat{\theta}-b)/\tilde{\sigma} \le \sqrt{n}(\hat{\theta}-\theta)/\tilde{\sigma} \le \sqrt{n}(\hat{\theta}-a)/\tilde{\sigma},$$

which leads to the following confidence interval of the (asymptotic) confidence level $1-\alpha$:

$$\left(\hat{\theta} \pm \tilde{\sigma} z_{\alpha/2}/\sqrt{n}\right)$$

For $\alpha = 0.05$, we recover the "placing a 2-standard deviation bound on the error".

In practice, we often do not know $\tilde{\sigma}$ and we have to approximate it with an estimator that converges to the true value.

**Thm 1.** *(Slutsky) Assume that the sequence of r.v.'s $\{X_n\}$ converges to $X$ in distribution and that $\{Z_n\}$ converges to a constant $c$ in probability. Then,*

- $X_n + Z_n \to X + c$,
- $X_n Z_n \to cX$,
- $X_n/Z_n \to X/c$ *provided* $c \neq 0$,

*where each convergence is in the distribution sense.*

Using the above theorem and assuming that $\sqrt{n}(\hat{\theta}-\theta) \to N(0,\tilde{\sigma}^2)$ and that $\hat{\sigma}^2$ is a consistent estimator of $\tilde{\sigma}^2$, we conclude that

$$\sqrt{n}(\hat{\theta}-\theta)/\hat{\sigma} \to N(0,\tilde{\sigma}^2)/\tilde{\sigma} = N(0,1)$$

Thus, we can use the pivot

$$G(Y_1,\dots,Y_n;\theta) := \sqrt{n}(\hat{\theta}-\theta)/\hat{\sigma}.$$

This, in fact, justifies the use of $S$ in place of $\sigma$ in the previous lectures, where we need to place a 2-standard deviation bound on the error.

**Ex 5.** *In this example, we aim to find the average time $\mu$ that a typical customer spends shopping in a store. We record the shopping times of $n = 64$ randomly chosen customers, and the realized values of the sample mean and sample variance are $\bar{y} = 33$, $s^2 = 256$.*

**Q 7.** *Find an asymptotic $0.9$-confidence interval for the true mean $\mu$.*

*We follow the method described above:*

$$\sqrt{n}(\bar{Y}-\mu)/S \to N(0,1),$$

$$\left(33 \pm 1.645\, s/\sqrt{64}\right) \approx \left(33 \pm 1.645\sqrt{256}/\sqrt{64}\right) = (29.71, 36.29).$$

The following theorem also helps.

**Thm 2.** *Assume that $X_n$ and $Z_n$ are independent and converge in distribution to $X$ and $Z$, respectively. Then, $X$ and $Z$ are independent and*

- $X_n + Z_n \to X + Z$,
- $X_n Z_n \to XZ$,
- $X_n/Z_n \to X/Z$ *provided* $Z \neq 0$ *with probability one.*

**Ex 6.** *A fridge of brand A fails during the warrantee period with (unknown) probability $p_1$. A fridge of brand B fails during the warrantee period with (unknown) probability $p_2$. Two samples (one of each brand) are collected.*
*A: sample of* $50$, $12$ *fail.*
*B: sample of* $60$, $12$ *fail.*

**Q 8.** *Provide a* $0.98$*-confidence interval for $p_1 - p_2$.*

*An unbiased estimator of $p_1 - p_2$ is $\bar{Y} - \bar{Z}$. Also,*

$$\sqrt{n_1}(\bar{Y}-p_1) \to N(0, p_1(1-p_1)), \quad \sqrt{n_2}(\bar{Z}-p_2) \to N(0, p_2(1-p_2))$$

*As the two samples are independent, we can use the above theorem to conclude that*

$$\sqrt{n_1}(\bar{Y}-p_1) - \sqrt{n_2}(\bar{Z}-p_2) \to N(0, p_1(1-p_1) + p_2(1-p_2)).$$

*If we assume that*

$$\frac{|n_1 - n_2|}{\sqrt{n_1} + \sqrt{n_2}}$$

*remains bounded for all $n_1, n_2$ as they increase to $\infty$, then*

$$(\sqrt{n_2} - \sqrt{n_1})(\bar{Z}-p_2) = \frac{n_2 - n_1}{\sqrt{n_1} + \sqrt{n_2}}(\bar{Z}-p_2) \to 0.$$

*Therefore,*

$$\sqrt{n_1}(\bar{Y} - \bar{Z} - (p_1 - p_2)) \to N(0, p_1(1-p_1) + p_2(1-p_2)),$$
$$\sqrt{n_2}(\bar{Y} - \bar{Z} - (p_1 - p_2)) \to N(0, p_1(1-p_1) + p_2(1-p_2)).$$

*Next, we need to find a consistent estimator $\hat{\sigma}^2$ for*

$$\tilde{\sigma}^2 = p_1(1-p_1) + p_2(1-p_2).$$

*As we do know that $\bar{Y}$ and $\bar{Z}$ are consistent estimators for $p_1$ and $p_2$, we naturally choose*

$$\hat{\sigma}^2 := \bar{Y}(1-\bar{Y}) + \bar{Z}(1-\bar{Z}) \to \tilde{\sigma}^2 = p_1(1-p_1) + p_2(1-p_2).$$

*Now, we can use the pivot*

$$G := \frac{\sqrt{n_1}(\bar{Y} - \bar{Z} - (p_1 - p_2))}{\hat{\sigma}} = \frac{\sqrt{n_1}(\bar{Y} - \bar{Z} - (p_1 - p_2))}{\sqrt{\bar{Y}(1-\bar{Y}) + \bar{Z}(1-\bar{Z})}} \to \frac{N(0, p_1(1-p_1) + p_2(1-p_2))}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}} = N(0,1),$$

*and repeat the general algorithm to obtain the following asymptotic conf. interval*

$$\left[\bar{Y} - \bar{Z} \pm z_{0.01}\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_1}\right].$$

*In fact, since $n_1$ and $n_2$ are not allowed to be too far away from each other, one can show that*

$$\frac{(\bar{Y} - \bar{Z} - (p_1 - p_2))}{\sqrt{\bar{Y}(1-\bar{Y})\frac{1}{n_2} + \bar{Z}(1-\bar{Z})\frac{1}{n_2}}} = \frac{\sqrt{n_2}(\bar{Y} - \bar{Z} - (p_1 - p_2))}{\sqrt{\bar{Y}(1-\bar{Y}) + \bar{Z}(1-\bar{Z})}} \to N(0,1),$$

$$\frac{(\bar{Y} - \bar{Z} - (p_1 - p_2))}{\sqrt{\bar{Y}(1-\bar{Y})\frac{1}{n_1} + \bar{Z}(1-\bar{Z})\frac{1}{n_2}}} \to N(0,1).$$

*Using the above as pivots, we obtain two other confidence intervals:*

$$\left[\bar{Y} - \bar{Z} \pm z_{0.01}\sqrt{p_1(1-p_1)/n_2 + p_2(1-p_2)/n_2}\right],$$

$$\left[\bar{Y} - \bar{Z} \pm z_{0.01}\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}\right].$$

*Picking the last one, for symmetry, we get*

$$\left(0.24 - 0.2 \pm 2.33\sqrt{0.24(1-0.24)/50 + 0.2(1-0.2)/60}\right) = (-0.1451,\ 0.2251)$$

*Note that we cannot reject $p_1 = p_2$ at this confidence level.*