

Statistics 431

Statistical Inference

Lecture IV

Mikhail Traskin

University of Pennsylvania
The Wharton School
Department of Statistics

September 17, 2007

These notes closely follow sections 7.2–7.4 of the *Probability and Statistics for Engineering and the Sciences* by Jay L. Devore, 7th edition.

Throughout this lecture we shall assume that we are given a dataset (sample) x_1, \dots, x_n , where x_i 's are numbers. We shall discuss the following topics.

1. Confidence interval for a population proportion: Wilson's procedure.
2. Sample size for a confidence interval of a population proportion.
3. Confidence bounds for a population proportion.
4. Small sample confidence intervals.
5. Prediction intervals.
6. Confidence interval for the variance of the normal distribution.
7. Terminology.

1 Confidence Interval for a Population Proportion: Wilson's procedure

Let X_1, \dots, X_n be i.i.d. X with $X \in \{0, 1\}$ and $\mathbf{P}(X = 1) = p$, $\mathbf{P}(X = 0) = q = 1 - p$. p is a population proportion. An estimate of the population proportion is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

with $E\hat{p} = p$ and $\sigma_{\hat{p}}^2 = p(1-p)/n = pq/n$. Assume $np \geq 10$ and $n(1-p) \geq 10$. Then from Central Limit Theorem follows that $\hat{p} \sim \mathcal{N}(p, p(1-p)/n)$ (in fact, approximately normal). Therefore we can write

$$\mathbf{P} \left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2} \right) \simeq 1 - \alpha.$$

Replacing $<$ signs under the probability with $=$ we solve equations as follows

$$\begin{aligned} -z_{\alpha/2} &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \\ z_{\alpha/2}^2 p(1-p)/n &= (\hat{p} - p)^2 \\ z_{\alpha/2}^2 p(1-p)/n &= \hat{p}^2 - 2p\hat{p} + p^2 \\ \hat{p}^2 - 2p\hat{p} + p^2 - z_{\alpha/2}^2 p(1-p)/n &= 0 \\ p^2(1 + z_{\alpha/2}^2/n) - (2\hat{p} + z_{\alpha/2}^2/n)p + \hat{p}^2 &= 0 \end{aligned}$$

Solution of the above quadratic equation is

$$p = \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n}.$$

Therefore confidence interval has a form

$$\left(\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n}, \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n} \right). \quad (1)$$

The procedure we just used to compute the confidence interval is called *Wilson's procedure*. If n is large then $z_{\alpha/2}^2/n$ is small and we can approximate the above expression with

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right). \quad (2)$$

Notice that the latter form coincides with the large sample confidence interval we computed in the previous lecture, since $\sqrt{\hat{p}(1-\hat{p})/n}$ is an estimate of $\sigma_{\hat{p}}$. An old rule of thumb says that this approximation works if $\hat{p}(1-\hat{p})n > 5$. It has simpler form, but since computers are easily available, Wilson's procedure is preferred. It is argued by various authors (see text for details) that interval of the form (1) works even if one of the conditions $pn \geq 10$ and $n(1-p) \geq 10$ does not hold.

2 Sample Size and Confidence Interval of a Population Proportion

We could proceed from interval (1), but for the purposes of determining sample size an approximate formula (2) will suffice. So width of the interval is

$$2w = 2z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Rearranging we get

$$n = \left\lceil z_{\alpha/2}^2 \frac{\hat{p}(1 - \hat{p})}{w^2} \right\rceil.$$

Since we don't know \hat{p} in advance, we are going to use worst-case estimate $\hat{p} = 1/2$, therefore

$$n = \left\lceil \frac{z_{\alpha/2}^2}{4w^2} \right\rceil.$$

If true p is close to $1/2$, then we are fine, if it is far from $1/2$, then we are wasting efforts gathering too big a sample.

3 Confidence Bounds for a Population Proportion

Just as in the previous lecture, we could compute confidence bounds. For example, the upper confidence bound

$$\mathbf{P} \left(p < \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) \simeq 1 - \alpha.$$

Another option is to use approach similar to the one we used to compute interval (1).

4 Small Sample Confidence Intervals

Even if sample size is small and σ is unknown, we can compute a confidence interval if we assume that population distribution is normal. This is possible, because

$$T = \frac{\bar{X} - \mu}{S\sqrt{n}}$$

has a t -distribution with $n-1$ degrees of freedom. Let t_{ν} denotes a density of the t distribution with ν degrees of freedom. Then

1. Each t_{ν} curve is bell-shaped and centered at 0.
2. Each t_{ν} curve is more spread out than the standard normal (z) curve.
3. As ν increases, the spread of the corresponding t_{ν} curve decreases.

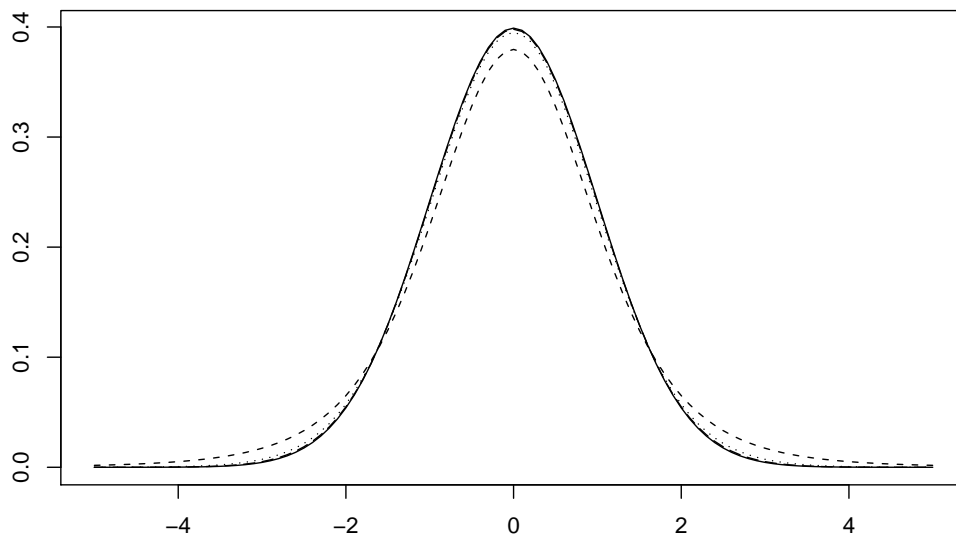


Figure 1: Various t -curves compared to the z curve (solid). Dashed line shows the t_5 curve, dotted line shows the t_{25} curve and long dashed line (almost coincides with the z -curve) corresponds to the t_{100} curve.

4. z curve is a t_ν curve with $\nu = \infty$.

Figure 1 illustrates several t -curves and compares them to the normal curve. Why is it that number of degrees of freedom of T is $n - 1$? Because we use n deviation $X_i - \bar{X}$ to compute S , but $\sum(X_i - \bar{X}) = 0$ implies that only $n - 1$ of these are freely determined. Define $t_{\alpha,\nu}$ as

$$\mathbf{P}(T_\nu > t_{\alpha,\nu}) = \alpha.$$

Then we can use T to obtain a confidence interval very much like we did it in case of a normal distribution and known σ :

$$\mathbf{P}\left(-t_{\alpha/2,n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1 - \alpha,$$

hence CI is

$$\left(\bar{X} - t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}\right).$$

Similarly, an upper confidence bound is given by

$$\bar{X} + t_{\alpha,n-1} \frac{S}{\sqrt{n}}$$

and lower is

$$\bar{X} - t_{\alpha,n-1} \frac{S}{\sqrt{n}}.$$

5 Prediction Intervals

Suppose after observing a sample X_1, \dots, X_n we want to predict a single new value X_{n+1} . Obvious prediction $\bar{X} = (\sum_{i=1}^n X_i)/n$ is good, but does not say anything about accuracy of the prediction. However, we can recall that \bar{X} and X_{n+1} are independent, normally distributed r.v.s, therefore $\bar{X} - X_{n+1}$ is also normally distributed and

$$E(\bar{X} - X_{n+1}) = 0$$

and

$$\text{Var}(\bar{X} - X_{n+1}) = \left(\frac{1}{n} + 1\right) \sigma^2.$$

Then

$$\mathbf{P} \left(-z_{\alpha/2} < \frac{\bar{X} - X_{n+1}}{\sigma \sqrt{\frac{1}{n} + 1}} < z_{\alpha/2} \right) = 1 - \alpha,$$

therefore the $100(1 - \alpha)\%$ prediction interval is

$$\left(\bar{X} - z_{\alpha/2} \sigma \sqrt{\frac{1}{n} + 1}, \bar{X} + z_{\alpha/2} \sigma \sqrt{\frac{1}{n} + 1} \right).$$

Since we usually do not know σ , we replace it by S . Then

$$\frac{\bar{X} - X_{n+1}}{S \sqrt{1 + \frac{1}{n}}}$$

has a t -distribution with $n - 1$ degrees of freedom and $100(1 - \alpha)\%$ prediction interval is

$$\left(\bar{X} - t_{\alpha/2, n-1} S \sqrt{\frac{1}{n} + 1}, \bar{X} + t_{\alpha/2, n-1} S \sqrt{\frac{1}{n} + 1} \right).$$

6 Confidence Interval for the Variance of the Normal Distribution

Since

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$$

has a χ_{n-1}^2 distribution and the above expression is clearly a pivot, then we can use it to compute a confidence interval for a population variance σ^2 :

$$\mathbf{P} \left(\chi_{1-\alpha/2, n-1}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2, n-1}^2 \right) = 1 - \alpha,$$

where $\chi_{\beta, \nu}^2$ is such that for $\gamma \sim \chi_{\nu}^2$

$$\mathbf{P}(\gamma > \chi_{\beta, \nu}^2) = \beta.$$

Rearranging we get $100(1 - \alpha)\%$ confidence interval for σ^2

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2}, \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2} \right).$$

A confidence interval for σ has lower and upper bounds that are square roots of the bounds in the interval for σ^2 .

7 Terminology

Standard deviation σ/\sqrt{n} of \bar{X} is often called *standard error*. If we estimate this SE by s/\sqrt{n} then it is called an *estimated standard error*.

Standard deviation of the estimator \hat{p} is $\sqrt{p(1-p)/n}$. It is also often called *standard error*. And if we use estimate \hat{p} instead of p , $\sqrt{\hat{p}(1-\hat{p})/n}$, then it is an *estimated standard error*.