

Statistics 431: Statistical Inference

Facts and Formulas

Probability foundations

The normal distribution and its samples

- The probability density function of a $N(\mu, \sigma^2)$ rv is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right).$$

- The population mean is μ ; the population SD is σ .
- For a sample X_1, \dots, X_n of size n from a normal population,
 - The sample estimate of μ is the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- The sample estimate of σ^2 is the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2 \right].$$

- The distribution of the normal sample mean: $\bar{X} \sim N(\mu, \sigma^2/n)$. The SD of \bar{X} , σ/\sqrt{n} , is also called the standard error (SE) of \bar{X} . We estimate it as S/\sqrt{n} .
- \bar{X} and S^2 are independent as rvs. The distribution of the sample variance: $(n-1)S^2 \sim \chi_{n-1}^2$.
- The sample histogram and the normal quantile plot are two graphical tools to judge whether a sample comes from an approximately normal population. Prefer the quantile plot.

The binomial distribution and its samples

- Let X count the number of “successes” in n independent Bernoulli trials, each with probability p of “success.” Then X has the binomial distribution, $X \sim \text{Bin}(n, p)$. The probability mass function of a binomial rv is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n.$$

- The population mean is np ; the population SD is $\sqrt{np(1-p)}$.
- For a binomial rv X based on a Bernoulli sample Z_1, \dots, Z_n from a population, the estimate of p is $\hat{p} = X/n$.
- The SD of \hat{p} is $\sqrt{p(1-p)/n}$; it is also called the SE of \hat{p} . We estimate it as $\sqrt{\hat{p}(1-\hat{p})/n}$.
- For large n , the distribution of \hat{p} is approximately $N(p, p(1-p)/n)$.

Chapter 7: Confidence intervals

One sample mean

- A $100\gamma\% = 100(1-\alpha)\%$ confidence interval (CI) for an unknown population mean μ has the general form

$$\bar{X} \pm C^* \cdot \frac{\sigma^*}{\sqrt{n}} = \left(\bar{X} - C^* \cdot \frac{\sigma^*}{\sqrt{n}}, \bar{X} + C^* \cdot \frac{\sigma^*}{\sqrt{n}} \right),$$

where C^* is an appropriate upper quantile and σ^* is an appropriate population SD or estimate thereof. The meaning of the confidence statement is that $P(\mu \in \text{Interval}) = \gamma$, at least approximately. The important situations are:

- σ known and either population normal or n large: $\sigma^* = \sigma$ and $C^* = z_{\alpha/2}$, the $(\alpha/2)$ upper quantile of the standard normal.
 - σ unknown and n large: $\sigma^* = S$ and $C^* = z_{\alpha/2}$.
 - σ unknown and population normal: $\sigma^* = S$ and $C^* = t_{\alpha/2; n-1}$, the $(\alpha/2)$ upper quantile of the t distribution with $n-1$ degrees of freedom (df).
- The sample size needed to get an interval of width w is (approximately)

$$n(w) = \left(2z_{\alpha/2} \frac{\sigma}{w} \right)^2,$$

rounded up to the nearest integer. When σ is unknown, use an estimate from previous experience or from the corresponding value of S in a pilot experiment.

One population proportion

- When n is large (say, $n\hat{p}(1-\hat{p}) > 20$) and \hat{p} is not too near 0 or 1, you can use the classical large-sample CI formula,

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

- Otherwise, use Wilson's formula

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

- When in doubt, use Wilson's formula – it is almost equivalent to the large-sample formula on large samples, but is better on small samples.
- The sample size needed to get an interval of width w is

$$n(w) = \left(2z_{\alpha/2} \frac{\sqrt{\tilde{p}(1 - \tilde{p})}}{w} \right)^2 ,$$

rounded up to the nearest integer. Here \tilde{p} is a prior guess for p . A worst-case (possibly too large) n is obtained by letting $\tilde{p} = 1/2$.

One-sided confidence confidence interval (aka confidence bound)

- A $100(1 - \alpha)\%$ upper confidence bound results from replacing \pm in the above by $+$, and $z_{\alpha/2}$ by z_{α} (or $t_{\alpha/2;n-1}$ by $t_{\alpha;n-1}$). The other end of the interval is $-\infty$.
- A $100(1 - \alpha)\%$ lower confidence bound results from replacing \pm in the above by $-$, and $z_{\alpha/2}$ by z_{α} (or $t_{\alpha/2;n-1}$ by $t_{\alpha;n-1}$). The other end of the interval is $+\infty$.

Prediction intervals

- Let Y be a single future observation from a normal population distribution. A $100(1 - \alpha)\%$ prediction interval for Y takes the form $\bar{X} \pm C^* \cdot \sigma^* \sqrt{1 + 1/n}$, with C^* and σ^* as above. The interval has the property that $P(Y \in \text{Interval}) \approx 1 - \alpha$.

Chapter 8: Hypothesis tests

General theory

- Tests involve a null hypothesis H_0 and an alternative hypothesis H_A . A typical case tests $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ for an unknown mean parameter μ . This is a two-sided test. A one-sided test looks like $H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$. To conduct the test, we need a test statistic T and a rejection region like $|T| > c$ or $T > c$. Here c is the critical value.
- Suppose we are testing $H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$. Then

$$P_{\mu \in H_0}(\text{Reject } H_0) = P(\text{Type I error}) \text{ for this } \mu ,$$

and

$$P_{\mu \in H_A}(\text{Do not reject } H_0) = P(\text{Type II error}) = \beta \text{ for this } \mu .$$

The significance level α is the probability of a Type I error at the boundary value μ_0 . The power of the test is $P_{\mu \in H_A}(\text{Reject } H_0) = 1 - \beta$.

- If we observe the test statistic value $T = t$, the p -value is the smallest α at which we can reject H_0 using t . If you know the p -value of a test, you know the outcome for *every* level α :

$$p\text{-value} < \alpha \Rightarrow \text{reject} ; p\text{-value} \geq \alpha \Rightarrow \text{do not reject} .$$

- Duality: for the usual two-sided tests, a level α test does *not* reject $H_0 : \mu = \mu_0$ exactly when a $100(1 - \alpha)\%$ CI for μ contains μ_0 . (There is a similar relationship between one-sided tests and upper/lower confidence bounds.)

Particular tests: one population mean

- A test of $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ rejects when $|T| > C^*$, where

$$T = \frac{\bar{X} - \mu_0}{\sigma^*/\sqrt{n}} .$$

Here σ^* and C^* are as in the above discussion of two-sided confidence intervals.

- The p -value corresponding to $T = t$ can be found by looking up $P(|T| > t)$ for the normal distribution (when n is large) or the t_{n-1} distribution (when n is small). NOTE: because of the absolute value signs, you must multiply the tabled value by 2.
- The sample size n at which a two-sided level α test has power $1 - \beta$ under the alternative μ' is approximately

$$n = \left(\frac{\sigma^*(z_{\alpha/2} + z_{\beta})}{\mu_0 - \mu'} \right)^2 .$$

In the one-sided case, put z_{α} in place of $z_{\alpha/2}$. (The resulting n is only valid if it is large, since the formula uses large-sample normality).

- A test of $H_0 : \mu \geq \mu_0$ vs $H_A : \mu < \mu_0$ would reject when $T < C^*$, where C^* is from the lower confidence bound case discussed above. p -values can be found analogously (do not multiply the tabled value by 2).

Particular tests: one population proportion

- To test $H_0 : p = p_0$ vs $H_A : p \neq p_0$, use

$$T = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} ,$$

with the critical value determined in the usual manner as a standard normal upper quantile. p -values are also determined from $T = t$ in an analogous manner. Here, n should not be too small; $np_0(1 - p_0) > 5$ should suffice. For smaller n , there is a procedure we have not covered based on the binomial distribution.

- The n for which a two-sided test of p_0 has power $1 - \beta$ under the alternative $p = p'$ is approximately

$$n = \left(\frac{z_{\alpha/2}\sqrt{p_0(1 - p_0)} + z_{\beta}\sqrt{p'(1 - p')}}{p' - p_0} \right)^2,$$

rounded up to the nearest integer. For a one-sided test, replace $z_{\alpha/2}$ with z_{α} .

Chapter 9: Inferences based on two samples

Inferences about the difference of two population means

- A two-sided hypothesis test has the form $H_0 : \mu_1 - \mu_2 = \Delta_0$ vs $H_A : \mu_1 - \mu_2 \neq \Delta_0$. Often, $\Delta_0 = 0$. If the sample from population A is independent of the sample from population B, reject when $|T| > C^*$, where

$$T = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{(\sigma_1^*)^2/n_1 + (\sigma_2^*)^2/n_2}}.$$

Here σ_1^* , σ_2^* , and C^* are like the values in the one-sample procedures. However, if n_1 or n_2 is small and σ is unknown, you need to assume normal population distributions and treat T as having a t distribution with ν df. Here

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}},$$

rounded to the nearest integer. Note that $\min\{n_1 - 1, n_2 - 1\} \leq \nu \leq n_1 + n_2 + 1$.

- A $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ takes the form $\bar{X} - \bar{Y} \pm C^* \cdot \text{SE}$, where SE is the denominator of the test statistic T .
- p -values can be found in the usual way from the value $T = t$ and the corresponding table.
- Formulas for β can be derived from the structure of the test. Samples size calculations are complicated, except in special cases. We do not give general formulas here, or for two proportions below.
- If the additional assumption $\sigma_1 = \sigma_2$ is tenable, use

$$T = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{S_{\text{pooled}}^2(1/n_1 + 1/n_2)}},$$

where

$$S_{\text{pooled}}^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 .$$

- When each observation from sample A is paired with an observation from sample B, inferences are based on the differences $D_i = X_i - Y_i$. The two-sample test of $H_0 : \mu_1 - \mu_2 = \Delta_0$ then reduces to a one-sample test of $H_0 : \mu_D = \Delta_0$, with the D_i 's as the sample.

Inferences about the difference of two population proportions

- When testing $H_0 : p_1 - p_2 = \Delta_0 \neq 0$, the situation is identical to the two-means case, except (\hat{p}_1, \hat{p}_2) replaces (\bar{X}, \bar{Y}) and $\hat{p}_i(1 - \hat{p}_i)$ replaces $(\sigma_i^*)^2$, $i = 1, 2$.
- When testing $H_0 : p_1 - p_2 = 0$, use

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}} ,$$

where \hat{p} is the combined sample estimate of p , defined by

$$\hat{p} = \frac{X + Y}{n_1 + n_2} = \frac{n_1}{n_1 + n_2} \hat{p}_1 + \frac{n_2}{n_1 + n_2} \hat{p}_2 .$$

Here again, n_1 and n_2 should not be small.

- NOTE: To build a confidence interval for $p_1 - p_2$, you should not use the pooled variance estimate based on \hat{p} .