Stat 427/527: Advanced Data Analysis I

Review of Chapters 1-4

Instructor Yan Lu

THE UNIVERSITY OF NEW MEXICO

Population and Sample

Goal: want to use the sample information to make inferences about the population and its parameters.

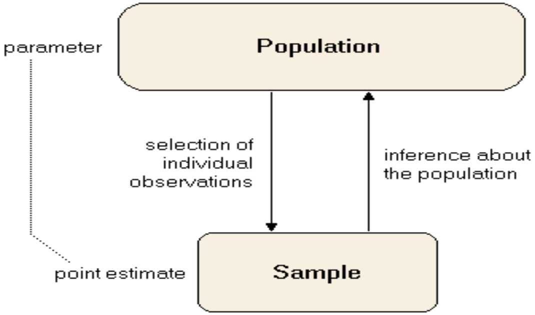

Figure 1 : Population, sample and statistical inference

Three basic assumptions

- Data are a random sample.
- The population frequency curve is normal.
- For the pooled variance two-sample test the population variances are also required to be equal.

Numerical and Graphical Summaries

Numerical summaries:

- measures of center (mean, median, mode)
- measures of spread (sample variance, sample standard deviation (SE), range, IQR)
  —Five numbers: minimum, Q1, Median, Q3, maximum

Graphical summaries:

- Stem and leaf plots
- Histograms
- Box Plots
  —-use boxplot to check for outliers
  —-use histogram and boxplot to describe the shape of the distribution. For example, skewed to left:
  Mean less than Median
  Median closer to Q3 than Q1, Median closer to max than min.
  Distance from min to Q1 greater than distance from Q3 to max.

- QQ plots
  —-assess the normality assumption
  —-The normality assumption is plausible if the plot is fairly linear.
  —-Rejection of normality assumption doesn't mean that the t-procedure inference is invalid. In fact, if there is no extreme outliers and no extreme skewness, t-procedure inference is usually valid.

Supplemental Tests:

- Normality

  —-Shapiro-Wilk test shapiro.test()

  —-Anderson-Darling test ad.test()

  —-Cramer-von Mises test cvm.test()

- Equal variance tests

  In the independent two sample t-test, we want to test

  $$H_0 : \sigma_1^2 = \sigma_2^2$$

  to decide between using the pooled-variance procedure or Satterthwaite's methods.

  ——suggest the pooled t-test and CI if H_0 is not rejected, and Satterthwaite's methods otherwise.

  Bartlett's test and Levene's test

  —Bartlett's test assumes the population distributions are normal

  —check normality prior to using Bartlett's test.

  —Levene's test is more robust to departures from normality than Bartlett's test; it is in the car package.

Inference for a population mean

Notations:

- Parameter of interest: population mean μ
- Sample mean: $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$.
- Observed sample mean: $\bar{y} = \sum_{i=1}^{n} y_i / n$

Two main methods for inferences on μ:

- **Confidence intervals** (CI)

  ——Construct CI based on different assumptions

  ——Interpret CI
- **Hypothesis tests**

  ——Construct the test statistic based on different assumptions

  ——Perform test, compare to critical value, or use p-value approach

  ——One sided, two sided

  ——Type I, II error, power of the test

Central limit theorem (CLT)

If $Y_1, \ldots, Y_n$ is a random sample of size n taken from a population or a distribution with mean μ and variance σ^2 and if $\bar{Y}$ is the sample mean, then for large n,

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

Standardization

If $Y_1, \ldots, Y_n$ is a random sample of size n taken from a normal population with mean μ and variance σ^2 and if $\bar{Y}$ is the sample mean, then, We may standardize $\bar{Y}$ by subtracting the mean and dividing by the standard deviation, which results in the variable

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

t distribution

t distribution is a continuous probability distribution that arises when estimating the mean of a normally distributed population in situations where the **sample size is small** and **population standard deviation is unknown**.

If $Z \sim N(0,1)$ and $V \sim \chi^2(v)$, and if Z and V independent, then the distribution of

$$T = \frac{Z}{\sqrt{V/v}}$$

is referred to as Student's t distribution with v degrees of freedom, denoted by $T \sim t(v)$.

If $Y_1, Y_2, \ldots, Y_n$ is a random sample from normal distribution with mean μ and variance σ^2 (σ^2 is unknown), i.e. $Y_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n.$ The r.v.

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

has a t distribution with $n-1$ degrees of freedom.
Proof: Let $Z = \sqrt{n}(\bar{Y} - \mu)/\sigma \sim N(0,1)$, and by $V = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, and by the fact that $\bar{Y}$ and S^2 are independent. $T = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom.

Sampling distribution

One sample problem:

- inference is based on the assumption that the random sample is from a normal population
—-therefore, sampling distribution of the mean $\bar{Y}$ is normal
—-and $\dfrac{\bar{Y}-\mu}{S/\sqrt{n}}$ is a t distribution with $n-1$ degrees of freedom.
- However, the t distribution based CI and hypothesis tests are relatively robust to the normality assumption.
—-Therefore, small to moderate departures from normality are not a cause of concern.
—-Remember that with the exponential distribution, which is highly skewed to the right, with sample size $n=6$, the bootstrapped $\bar{Y}$ is close to normal.
—-As long as no extreme skewness and no extreme outliers, violation of normality are not a cause of concern.

Two sample problem:

- inferece is based on the assumption that the two independent random samples are from two independent normal population —-therefore, sampling distribution of the difference in means, $d = \bar{Y}_1 - \bar{Y}_2$, is normal.
  — and $\dfrac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{SE(\bar{Y}_1 - \bar{Y}_2)}$ is a t distribution with certain degrees of freedom.
  —-$SE(\bar{Y}_1 - \bar{Y}_2)$ have different forms with equal variance and unequal variance assumptions.
- However, the t distribution based CI and hypothesis tests are relatively robust to the normality assumption.
  —-Therefore, small to moderate departures from normality are not a cause of concern.
  —-Remember that with the exponential distribution, which is highly skewed to the right, with sample size $n = 6$, the bootstrapped $\bar{Y}$ is close to normal.
  —-As long as no extreme skewness and no extreme outliers, violation of normality are not a cause of concern.

Confidence Intervals

Table 1 : Confidence Intervals for μ

Popn distribution Size Popn sd σ	Normal Any Known	Any $n \geq 40$ (large sample) Unknown	Normal $n < 40$ Unknown
Parameter Estimate SE of the estimator	μ $\bar{y}$ $\frac{\sigma}{\sqrt{n}}$	μ $\bar{y}$ $\frac{s}{\sqrt{n}}$	μ $\bar{y}$ $\frac{s}{\sqrt{n}}$
Distribution CI	$\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ $\bar{y} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$	$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim N(0,1)$ $\bar{y} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ $\bar{y} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$

Hypothesis Tests

Table 2 : $\bar{y}$ is sample mean and s is sample standard deviation; $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ percentage points of the t distribution with $n-1$ degrees of freedom; $t_{n-1,\alpha}$ is the upper α percentage points of the t distribution with $n-1$ degrees of freedom; T_{n-1} is a random variable following t distribution with $n-1$ degrees of freedom. α is the significance level of the test

Step 1:	$H_1 : \mu \neq \mu_0$
Step 2:	compute $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$
Step 3a:	Reject H_0 if $t_0 > t_{n-1,\alpha/2}$ or $t_0 < -t_{n-1,\alpha/2}$
Step 3b:	P-value $=2P(T_{n-1} > \|t_0\|)$ Reject H_0 if P-value $< \alpha$
Power	$P(T_0 > t_{n-1,\alpha/2}\|\mu_1) + P(T_0 < -t_{n-1,\alpha/2}\|\mu_1)$

Table 3 : $\bar{y}$ is sample mean and s is sample standard deviation; $t_{n-1,\alpha/2}$ is the upper $\alpha/2$ percentage points of the t distribution with $n-1$ degrees of freedom; $t_{n-1,\alpha}$ is the upper α percentage points of the t distribution with $n-1$ degrees of freedom; T_{n-1} is a random variable following t distribution with $n-1$ degrees of freedom. α is the significance level of the test

Step 1:	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$
Step 2:	compute $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$	compute $t_0 = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$
Step 3a:	Reject H_0 if $t_0 < -t_{n-1,\alpha}$	Reject H_0 if $t_0 > t_{n-1,\alpha}$
Step 3b:	P-value $=P(T_{n-1} < t_0)$ Reject H_0 if P-value $< \alpha$	P-value $=P(T_{n-1} > t_0)$ Reject H_0 if P-value $< \alpha$
Power	$P(T_0 < -t_{n-1,\alpha/2} \mid \mu_1)$	$P(T_0 > t_{n-1,\alpha/2} \mid \mu_1)$

Paired Versus Independent Samples

Suppose you have two populations of interest, say populations 1 and 2

- Interested in comparing their (unknown) population means, μ_1 and μ_2.
- Inferences on the unknown population means are based on samples from each population. In practice, most problems fall into one of two categories.

  Independent samples where the sample taken from population 1 has no effect on which observations are selected from population 2, and vice versa.

  Paired or dependent samples where experimental units are paired based on factors related or unrelated to the variable measured. Note that with paired data, the sample sizes are equal to the number of pairs.

Confidence Intervals and Hypothesis Tests

Confidence Intervals and Tests for two independent sample problem:

- Assume that the two independent populations have normal frequency curves with variances unknown.
- Let $(n_1,\ \bar{y}_1,\ s_1)$ and $(n_2,\ \bar{y}_2,\ s_2)$ be the sample sizes, means and standard deviations from the two samples.

Table 4 : CI and Tests for two independent sample problem

	Pooled, Equal Variances	Satterthwaite's, unequal variances
Parameters	$\mu_1 - \mu_2$	$\mu_1 - \mu_2$
Estimate	$\bar{y}_1 - \bar{y}_2$	$\bar{y}_1 - \bar{y}_2$
$SE_{\bar{Y}_1 - \bar{Y}_2}$	$s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $s^2_{\text{pooled}} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
df	$df1 = n_1 + n_2 - 2$	$df2 = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$
Distribution	$\dfrac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{df1}$	$\dfrac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{df2}$
CI	$(\bar{y}_1 - \bar{y}_2) \pm t_{\text{crit, df}} SE_{\bar{Y}_1 - \bar{Y}_2}$	
Hypothesis	$H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$	
Test Statistics	$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}$	
Reject H_0	if $\|t_s\| > t_{crit,df}$	