Lesson - Statistical inference

Year: University Year 1

Statistical inference uses a sample to learn about a population. This chapter covers sampling distributions, confidence intervals, hypothesis tests, chi-squared tests and regression.

Starter activity

A juice company claims that its cartons contain 1 litre. You measure 10 cartons and find a mean of 0.985 litres. Is the company cheating, or could this happen by chance? What extra information would you need to decide? Discuss in pairs and write down how you would make the decision fair to both the company and the customers.

Samples and populations

Definition: parameter and statistic

A **parameter** (such as the population mean μ or proportion p) is fixed but unknown. A **statistic** (such as the sample mean $\bar{x}$) is computed from data and varies from sample to sample.

Rule: sample mean and variance

$\bar{x} = \frac{1}{n}\sum x_i$ and the **sample variance** is $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$. Dividing by $n-1$ makes s^2 an unbiased estimator of σ^2.

Example 1

Find $\bar{x}$ and s^2 for $4, 7, 6, 9, 4$.

$\bar{x} = 6$ and $s^2 = \frac{4+1+0+9+4}{4} = 4.5$.

Sampling distribution and the central limit theorem

Property: distribution of the sample mean

If $X_1, \dots, X_n$ are independent with mean μ and variance σ^2, then $E[\bar{X}] = \mu$ and $\operatorname{Var} \bar{X} = \frac{\sigma^2}{n}$. The **standard error** is $\frac{\sigma}{\sqrt{n}}$; for example $\frac{12}{\sqrt{36}} = 2$.

Theorem: central limit theorem

For large n (typically $n \geq 30$), $\bar{X}$ is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$ whatever the shape of the

population. If the population is normal, $\bar{X}$ is exactly normal.

Confidence intervals

Rule: confidence intervals

A 95% confidence interval for μ with known σ is $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. With σ unknown and a normal population, use s and the t-value with $n-1$ degrees of freedom. For a proportion, $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. To achieve margin of error E take $n \geq \left(\frac{1.96\sigma}{E}\right)^2$.

Example 2

Find 95% intervals for (a) $\bar{x} = 50$, $\sigma = 12$, $n = 36$; (b) $n = 16$, $\bar{x} = 20$, $s = 4$, $t_{15} = 2.131$.

(a) $50 \pm 1.96 \times 2 = (46.08,\ 53.92)$. (b) $20 \pm 2.131 \times 1 = (17.87, 22.13)$. For $\hat{p} = \frac{120}{300} = 0.4$ the margin is $1.96 \times 0.0283 = 0.0554$. With $\sigma = 10$ and $E = 1$, $n \geq 385$; halving the margin needs four times the sample size.

Watch out

The 95% means that about 95% of intervals built this way would contain μ if we repeated the sampling. It does *not* mean that μ lies in this one interval with probability 0.95.

Hypothesis testing

Definition: hypothesis test

A test compares a **null hypothesis** H_0 (a precise claim such as $\mu = 100$) with an **alternative** H_1 (one-tailed or two-tailed). The **test statistic** measures how far the data are from H_0; the **p-value** is the probability, assuming H_0 true, of a result at least as extreme as observed. Reject H_0 if the p-value is below the **significance level** α (commonly 0.05).

Definition: errors and power

A **Type I error** is rejecting a true H_0 (probability α). A **Type II error** is failing to reject a false H_0. The **power** is $1 - P(\text{Type II})$.

Method: carrying out a test

1. State H_0 and H_1.

2. Choose α and the test, and compute the test statistic.
3. Find the p-value or compare with the critical value.
4. Conclude in the context of the question.

Property: common tests

- **z-test (known σ):** $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.
- **t-test (unknown σ):** $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ with $n - 1$ degrees of freedom; a paired t-test applies this to the differences.
- **Two-sample z-test:** $z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$.
- **Proportion test:** $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.

Example 3: z-test

$H_0 : \mu = 100$, $\sigma = 15$, $n = 25$, $\bar{x} = 106$. Test at 5%.

$z = \frac{6}{3} = 2$. The two-tailed p-value is 0.0455, so reject H_0. (One-tailed p-value: 0.0228.)

Example 4: t-test and binomial test

(a) $n = 9$, $\bar{x} = 52$, $s = 6$, $\mu_0 = 50$. (b) 9 heads in 10 tosses, $H_0 : p = 0.5$, $H_1 : p > 0.5$.

(a) $t = 1$, critical value 2.306 (8 d.f., 5% two-tailed): do not reject. (b) $P(X \geq 9) = \frac{11}{1024} = 0.0107$, so reject H_0.

Chi-squared tests

Rule: chi-squared statistic

$\chi^2 = \sum \frac{(O-E)^2}{E}$. In a **goodness-of-fit** test the degrees of freedom are (categories -1) minus fitted parameters. In a **test of independence** on an $r \times c$ table, $E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$ and d.f. $= (r-1)(c-1)$.

Example 5

A die rolled 60 times gives counts $8, 12, 9, 11, 10, 10$. Is it biased?

$\chi^2 = \frac{4+4+1+1+0+0}{10} = 1.0$, with 5 d.f. and critical value 11.07. There is no evidence of bias.

Correlation and regression

Rule: least-squares line

The least-squares line $y = a + bx$ has $b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$. The correlation coefficient is $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, where $S_{xy} = \sum(x - \bar{x})(y - \bar{y})$. Predict only within the range of the data (interpolation).

Watch out

- Using n instead of $n - 1$ in s^2.
- Saying a large p-value proves H_0.
- Claiming correlation implies causation.

Inference in practice

Key fact: z or t?

When the population standard deviation σ is known, or when the sample is large (say $n \geq 30$), the standardised mean $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is treated as $N(0, 1)$. When σ is unknown and estimated by s from a small sample of a normal population, the extra uncertainty is accounted for by Student's t distribution with $n - 1$ degrees of freedom. The t distribution has heavier tails than the normal, so its critical values are larger: $t_{5,\,0.975} = 2.571$ against $z = 1.96$. As n grows, t_{n-1} approaches $N(0, 1)$.

Method: writing up a hypothesis test

1. Define the parameter in context and state H_0 and H_1, deciding between one-sided and two-sided before looking at the data.
2. Choose the significance level α and the test, and state its assumptions.
3. Compute the test statistic from the data.
4. Find the p-value or compare with the critical value.
5. Decide: reject H_0 if p.
6. Conclude in context, using cautious language: "there is evidence that..." or "there is not enough evidence that...".

Worked example 6: A confidence interval for a difference of means

Two large classes take the same test. Class 1: $n = 60$, $\bar{x} = 64$, $s = 12$. Class 2: $n = 50$, $\bar{x} = 59$, $s = 10$. Find a 95% confidence interval for $\mu_1 - \mu_2$.

The standard error of the difference is $\sqrt{\frac{144}{60}+\frac{100}{50}}=\sqrt{2.4+2}=\sqrt{4.4}\approx 2.10$. The interval is $5\pm 1.96\times 2.10=5\pm 4.11$, that is $(0.89,\ 9.11)$. Since the interval does not contain 0, a two-sided test at the 5% level would reject $H_0:\mu_1=\mu_2$. Intervals and tests carry the same information, but the interval also shows how large the difference might be.

Worked example 7: A binomial test for a small sample

A coin gives 9 heads in 12 tosses. Is there evidence at the 5% level that it is biased towards heads?

Let $X\sim B(12,0.5)$ under $H_0:p=0.5$, with $H_1:p>0.5$. The p-value is $P(X\geq 9)=\frac{\binom{12}{9}+\binom{12}{10}+\binom{12}{11}+\binom{12}{12}}{4096}=\frac{220+66+12+1}{4096}=\frac{299}{4096}\approx 0.073$. Since $0.073>0.05$, we do not reject H_0. With small samples it is better to use the exact binomial distribution than a normal approximation.

Worked example 8: Choosing a sample size for a proportion

A council wants to estimate the proportion of households that recycle glass, to within 3 percentage points with 95% confidence. How many households should it survey?

The margin of error is $1.96\sqrt{\frac{p(1-p)}{n}}$. Without prior information we use the worst case $p=0.5$, which maximises $p(1-p)$. Then $1.96\sqrt{\frac{0.25}{n}}\leq 0.03$ gives $n\geq\left(\frac{1.96\times 0.5}{0.03}\right)^2\approx 1067.1$, so 1 068 households. Halving the margin of error would need four times as many.

Worked example 9: Interpreting a regression output

A regression of exam mark y on revision hours x for 25 students gives $y=38+2.4x$ with $r=0.62$. Interpret the result.

The gradient means that each extra hour of revision is associated with about 2.4 more marks, and the intercept 38 is the predicted mark with no revision (meaningful only if $x=0$ lies within the data). The value $r^2=0.38$ says that about 38% of the variation in marks is explained by the linear relationship with revision time. The test of $\rho=0$ gives $t=0.62\sqrt{23/(1-0.3844)}\approx 3.79$, highly significant, but the data are observational, so other factors such as motivation may explain part of the link.

Worked example 10: A goodness-of-fit test with an estimated parameter

The numbers of goals in 100 school football matches give a mean of 1.5. A Poisson model is fitted. How many degrees of freedom does the χ^2 test have if the data are grouped as 0, 1, 2, 3 and "4 or more" goals?

There are 5 classes, and the totals must agree, which removes one degree of freedom. The mean of the Poisson distribution was estimated from the data, which removes another. So the test has $5 - 1 - 1 = 3$ degrees of freedom, and the 5% critical value is 7.815. The expected frequencies are $100\,P(X = k)$ for $X \sim \mathrm{Po}(1.5)$: 22.3, 33.5, 25.1, 12.6 and 6.6. All are above 5, so no classes need to be merged before calculating χ^2.

Note: p-values and statistical significance

A p-value is the probability, assuming H_0 is true, of getting a result at least as extreme as the one observed. It is not the probability that H_0 is true. A very small effect can be statistically significant with a huge sample, and an important effect can fail to be significant with a small one. Good reports therefore give an estimate with a confidence interval as well as a p-value, and they state in advance which tests will be done, to avoid searching through many tests until one happens to be significant.

Watch out: typical mistakes

- Saying "we accept H_0" or "H_0 is proved": we only fail to reject it.
- Using z instead of t for a small sample with unknown σ.
- Using a χ^2 test with expected frequencies below 5: merge classes first.
- Extrapolating a regression line or reading causation into a correlation.

Did you know?

William Sealy Gosset, who worked for the Guinness brewery in Dublin, published the t distribution in 1908 under the pen name "Student", because his employer did not allow staff to publish under their own names. Ronald Fisher popularised significance testing in the 1920s, and Jerzy Neyman and Egon Pearson introduced Type I and Type II errors and power, ideas used today in every clinical trial.

Otto's tip

Always state H_0, H_1, the statistic, the critical value or p-value, and a conclusion in the context of the question.