

Sampling and Hypothesis Testing

Allin Cottrell

Population and sample

Population: an entire set of objects or units of observation of one sort or another.

Sample: subset of a population.

Parameter versus *statistic*.

	size	mean	variance	proportion
Population:	N	μ	σ^2	π
Sample:	n	\bar{x}	s^2	p

1

Properties of estimators: sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

To make inferences regarding the population mean, μ , we need to know something about the probability distribution of this sample statistic, \bar{x} .

The distribution of a sample statistic is known as a *sampling distribution*. Two of its characteristics are of particular interest, the mean or expected value and the variance or standard deviation.

$E(\bar{x})$: Thought experiment: Sample repeatedly from the given population, each time recording the sample mean, and take the average of those sample means.

2

If the sampling procedure is *unbiased*, deviations of \bar{x} from μ in the upward and downward directions should be equally likely; on average, they should cancel out.

$$E(\bar{x}) = \mu = E(X)$$

The sample mean is then an *unbiased estimator* of the population mean.

3

Efficiency

One estimator is more *efficient* than another if its values are more tightly clustered around its expected value.

E.g. alternative estimators for the population mean: \bar{x} versus the average of the largest and smallest values in the sample.

The degree of dispersion of an estimator is generally measured by the standard deviation of its probability distribution (sampling distribution). This goes under the name *standard error*.

4

Standard error of sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- The more widely dispersed are the population values around their mean (larger σ), the greater the scope for sampling error (i.e. drawing by chance an unrepresentative sample whose mean differs substantially from μ).
- A larger sample size (greater n) narrows the dispersion of \bar{x} .

5

Other statistics

Population *proportion*, π .

The corresponding sample statistic is the proportion of the sample having the characteristic in question, p .

The sample proportion is an unbiased estimator of the population proportion

$$E(p) = \pi$$

Its standard error is given by

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

6

Population *variance*, σ^2 .

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Estimator, sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

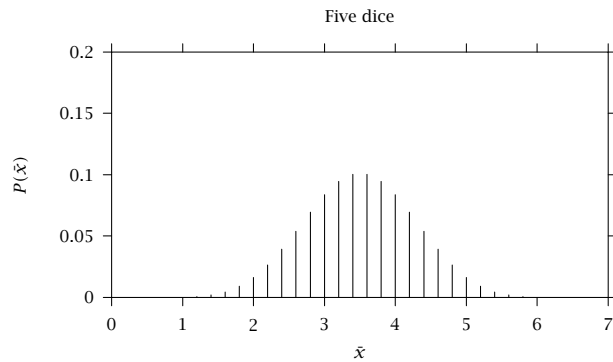
7

Shape of sampling distributions

Besides knowing expected value and standard error, we also need to know the *shape* of a sampling distribution in order to put it to use.

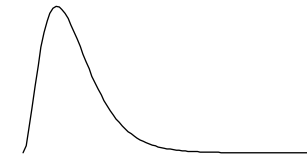
Sample mean: Central Limit Theorem implies a Gaussian distribution, for “large enough” samples.

Reminder:



8

Not all sampling distributions are Gaussian, e.g. sample variance as estimator of population variance. In this case the ratio $(n-1)s^2/\sigma^2$ follows a skewed distribution known as χ^2 , (*chi-square*) with $n-1$ degrees of freedom.



If the sample size is large the χ^2 distribution converges towards the normal.

9

Confidence intervals

If we know the mean, standard error and shape of the distribution of a given sample statistic, we can then make definite probability statements about the statistic.

Example: $\mu = 100$ and $\sigma = 12$ for a certain population, and we draw a sample with $n = 36$ from that population.

The standard error of \bar{x} is $\sigma/\sqrt{n} = 12/6 = 2$, and a sample size of 36 is large enough to justify the assumption of a Gaussian sampling distribution. We know that the range $\mu \pm 2\sigma$ encloses the central 95 percent of a normal distribution, so we can state

$$P(96 < \bar{x} < 104) \approx .95$$

There's a 95 percent probability that the sample mean lies within 4 units (= 2 standard errors) of the population mean, 100.

10

Population mean unknown

If μ is unknown we can still say

$$P(\mu - 4 < \bar{x} < \mu + 4) \approx .95$$

With probability .95 the sample mean will be drawn from within 4 units of the unknown population mean.

We go ahead and draw the sample, and calculate a sample mean of (say) 97. If there's a probability of .95 that our \bar{x} came from within 4 units of μ , we can turn that around: we're entitled to be 95 percent confident that μ lies between 93 and 101.

We draw up a 95 percent *confidence interval* for the population mean as $\bar{x} \pm 2\sigma_{\bar{x}}$.

11

Population variance unknown

With σ unknown, we have to *estimate* the standard error of \bar{x} .

$$s_{\bar{x}} \equiv \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

We can now reformulate our 95 percent confidence interval for μ :
 $\bar{x} \pm 2s_{\bar{x}}$.

Strictly speaking, the substitution of s for σ alters the shape of the sampling distribution. Instead of being Gaussian it now follows the t distribution, which looks very much like the Gaussian except that it's a bit "fatter in the tails".

12

The t distribution

Unlike the Gaussian, the t distribution is not fully characterized by its mean and standard deviation: there is an additional factor, namely the *degrees of freedom* (df).

- For estimating a population mean the df term is the sample size minus 1.
- At low degrees of freedom the t distribution is noticeably more "dispersed" than the Gaussian, meaning that a 95 percent confidence interval would have to be wider (greater uncertainty).
- As the degrees of freedom increase, the t distribution converges towards the Gaussian.
- Values enclosing the central 95 percent of the distribution:

$$\text{Normal: } \mu \pm 1.960\sigma$$

$$t(30): \mu \pm 2.042\sigma$$

13

Further examples

The following information regarding the Gaussian distribution enables you to construct a 99 percent confidence interval.

$$P(\mu - 2.58\sigma < x < \mu + 2.58\sigma) \approx 0.99$$

Thus the 99 percent interval is $\bar{x} \pm 2.58\sigma_{\bar{x}}$.

If we want greater confidence that our interval straddles the unknown parameter value (99 percent versus 95 percent) then our interval must be wider (± 2.58 standard errors versus ± 2 standard errors).

14

Estimating a proportion

An opinion polling agency questions a sample of 1200 people to assess the degree of support for candidate X.

- Sample info: $p = 0.56$.
- Our single best guess at the population proportion, π , is then 0.56, but we can quantify our uncertainty.
- The standard error of p is $\sqrt{\pi(1 - \pi)/n}$. The value of π is unknown but we can substitute p or, to be conservative, we can put $\pi = 0.5$ which maximizes the value of $\pi(1 - \pi)$.
- On the latter procedure, the estimated standard error is $\sqrt{0.25/1200} = 0.0144$.
- The large sample justifies the Gaussian assumption for the sampling distribution; the 95 percent confidence interval is $0.56 \pm 2 \times 0.0144 = 0.56 \pm 0.0289$.

15

Generalizing the idea

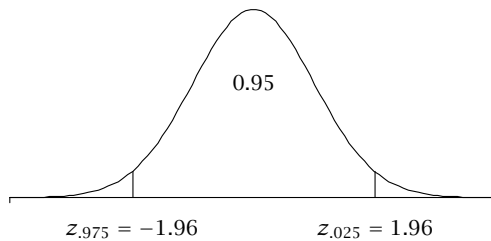
Let θ denote a “generic parameter”.

1. Find an estimator (preferably unbiased) for θ .
2. Generate $\hat{\theta}$ (point estimate).
3. Set confidence level, $1 - \alpha$.
4. Form interval estimate (assuming symmetrical distribution):

$$\hat{\theta} \pm \text{maximum error for } (1 - \alpha) \text{ confidence}$$

“Maximum error” equals so many standard errors of such and such a size. The number of standard errors depends on the chosen confidence level (possibly also the degrees of freedom). The size of the standard error, $\sigma_{\hat{\theta}}$, depends on the nature of the parameter being estimated and the sample size.

16



This is about as far as we can go in general terms. The specific formula for $\sigma_{\hat{\theta}}$ depends on the parameter.

18

Z-scores

Suppose the sampling distribution of $\hat{\theta}$ is Gaussian. The following notation is useful:

$$z = \frac{x - \mu}{\sigma}$$

The “standard normal score” or “z-score” expresses the value of a variable in terms of its distance from the mean, measured in standard deviations.

Example: $\mu = 1000$ and $\sigma = 50$. The value $x = 850$ has a z-score of -3.0 : it lies 3 standard deviations below the mean.

Where the distribution of $\hat{\theta}$ is Gaussian we can write the $1 - \alpha$ confidence interval for θ as

$$\hat{\theta} \pm \sigma_{\hat{\theta}} z_{\alpha/2}$$

17

The logic of hypothesis testing

Analogy between the set-up of a hypothesis test and a court of law.

Defendant on trial in the statistical court is the *null hypothesis*, some definite claim regarding a parameter of interest.

Just as the defendant is presumed innocent until proved guilty, the null hypothesis (H_0) is assumed true (at least for the sake of argument) until the evidence goes against it.

H_0 is in fact:

		H_0 is in fact:	
Decision:		True	False
Reject		Type I error $P = \alpha$	Correct decision
Fail to reject		Correct decision	Type II error $P = \beta$

$1 - \beta$ is the *power* of a test; trade-off between α and β .

19

Choosing the significance level

How do we get to *choose* α (probability of Type I error)?

The calculations that compose a hypothesis test are condensed in a key number, namely a conditional probability: *the probability of observing the given sample data, on the assumption that the null hypothesis is true.*

This is called the *p-value*. If it is small, we can place one of two interpretations on the situation:

- (a) The null hypothesis is true and the sample we drew is an improbable, unrepresentative one.
- (b) The null hypothesis is false.

The smaller the p-value, the less comfortable we are with alternative

(a). (Digression) To reach a conclusion we must specify the limit of our comfort zone, a p-value below which we'll reject H_0 .

20

Say we use a cutoff of .01: we'll reject the null hypothesis if the p-value for the test is $\leq .01$.

If the null hypothesis is in fact true, what is the probability of our rejecting it? It's the probability of getting a p-value less than or equal to .01, which is (by definition) .01.

In selecting our cutoff we selected α , the probability of Type I error.

21

Example of hypothesis test

A maker of RAM chips claims an average access time of 60 nanoseconds (ns) for the chips. Quality control has the job of checking that the production process is maintaining acceptable access speed: they test a sample of chips each day.

Today's sample information is that with 100 chips tested, the mean access time is 63 ns with a standard deviation of 2 ns. Is this an acceptable result?

Should we go with the symmetrical hypotheses

$$H_0: \mu = 60 \text{ versus } H_1: \mu \neq 60?$$

Well, we don't mind if the chips are faster than advertised.

So instead we adopt the asymmetrical hypotheses:

$$H_0: \mu \leq 60 \text{ versus } H_1: \mu > 60$$

Let $\alpha = 0.05$.

22

The p-value is

$$P(\bar{x} \geq 63 \mid \mu \leq 60)$$

where $n = 100$ and $s = 2$.

- If the null hypothesis is true, $E(\bar{x})$ is no greater than 60.
- The estimated standard error of \bar{x} is $s/\sqrt{n} = 2/10 = .2$.
- With $n = 100$ we can take the sampling distribution to be normal.
- With a Gaussian sampling distribution the *test statistic* is the z-score.

$$z = \frac{\bar{x} - \mu_{H_0}}{s_{\bar{x}}} = \frac{63 - 60}{.2} = 15$$

23

Variations on the example

Suppose the test were as described above, except that the sample was of size 10 instead of 100.

Given the small sample and the fact that the population standard deviation, σ , is unknown, we could not justify the assumption of a Gaussian sampling distribution for \bar{x} . Rather, we'd have to use the t distribution with $df = 9$.

The estimated standard error, $s_{\bar{x}} = 2/\sqrt{10} = 0.632$, and the test statistic is

$$t(9) = \frac{\bar{x} - \mu_{H_0}}{s_{\bar{x}}} = \frac{63 - 60}{.632} = 4.74$$

The p-value for this statistic is 0.000529—a lot larger than for $z = 15$, but still much smaller than the chosen significance level of 5 percent, so we still reject the null hypothesis.

24

In general the test statistic can be written as

$$\text{test} = \frac{\hat{\theta} - \theta_{H_0}}{s_{\hat{\theta}}}$$

That is, sample statistic minus the value stated in the null hypothesis—which by assumption equals $E(\hat{\theta})$ —divided by the (estimated) standard error of $\hat{\theta}$.

The distribution to which “test” must be referred, in order to obtain the p-value, depends on the situation.

25

Another variation

We chose an asymmetrical test setup above. What difference would it make if we went with the symmetrical version,

$$H_0: \mu = 60 \quad \text{versus} \quad H_1: \mu \neq 60?$$

We have to think: *what sort of values of the test statistic should count against the null hypothesis?*

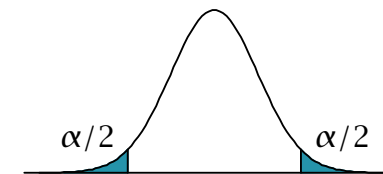
In the asymmetrical case only values of \bar{x} greater than 60 counted against H_0 . A sample mean of (say) 57 would be consistent with $\mu \leq 60$; it is not even *prima facie* evidence against the null.

Therefore the *critical region* of the sampling distribution (the region containing values that would cause us to reject the null) lies strictly in the upper tail.

But if the null hypothesis were $\mu = 60$, then values of \bar{x} both substantially below and substantially above 60 would count against it. The critical region would be divided into two portions, one in each tail of the sampling distribution.

26

$H_0: \mu = 60$. Two-tailed test. Both high and low values count against H_0 .



$H_0: \mu \leq 60$. One-tailed test. Only high values count against H_0 .



27

Practical consequence

We must double the p -value, before comparing it to α .

- The sample mean was 63, and the p -value was defined as the probability of drawing a sample “like this or worse”, from the standpoint of H_0 .
- In the symmetrical case, “like this or worse” means “with a sample mean this far away from the hypothesized population mean, or farther, in either direction”.
- So the p -value is $P(\bar{x} \geq 63 \cup \bar{x} \leq 57)$, which is double the value we found previously.

28

More on p -values

Let E denote the sample evidence and H denote the null hypothesis that is “on trial”. The p -value can then be expressed as $P(E|H)$.

This may seem awkward. Wouldn't it be better to calculate the conditional probability the other way round, $P(H|E)$?

Instead of working with the probability of obtaining a sample like the one we in fact obtained, assuming the null hypothesis to be true, why can't we think in terms of the probability that the null hypothesis is true, given the sample evidence?

29

Recall the multiplication rule for probabilities, which we wrote as

$$P(A \cap B) = P(A) \times P(B|A)$$

Swapping the positions of A and B we can equally well write

$$P(B \cap A) = P(B) \times P(A|B)$$

And taking these two equations together we can infer that

$$P(A) \times P(B|A) = P(B) \times P(A|B)$$

or

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)}$$

This is *Bayes' rule*. It provides a means of converting from a conditional probability one way round to the inverse conditional probability.

30

Substituting E (Evidence) and H (null Hypothesis) for A and B , we get

$$P(H|E) = \frac{P(H) \times P(E|H)}{P(E)}$$

We know how to find the p -value, $P(E|H)$. To obtain the probability we're now canvassing as an alternative, $P(H|E)$, we have to supply in addition $P(H)$ and $P(E)$.

$P(H)$ is the marginal probability of the null hypothesis and $P(E)$ is the marginal probability of the sample evidence.

Where are these going to come from??

31

Confidence intervals and tests

The symbol α is used for both the significance level of a hypothesis test (the probability of Type I error), and in denoting the confidence level $(1 - \alpha)$ for interval estimation.

There is an equivalence between a two-tailed hypothesis test at significance level α and an interval estimate using confidence level $1 - \alpha$.

Suppose μ is unknown and a sample of size 64 yields $\bar{x} = 50$, $s = 10$. The 95 percent confidence interval for μ is then

$$50 \pm 1.96 \left(\frac{10}{\sqrt{64}} \right) = 50 \pm 2.45 = 47.55 \text{ to } 52.45$$

Suppose we want to test $H_0: \mu = 55$ using the 5 percent significance level. No additional calculation is needed. The value 55 lies outside of the 95 percent confidence interval, so we can conclude that H_0 is rejected.

32

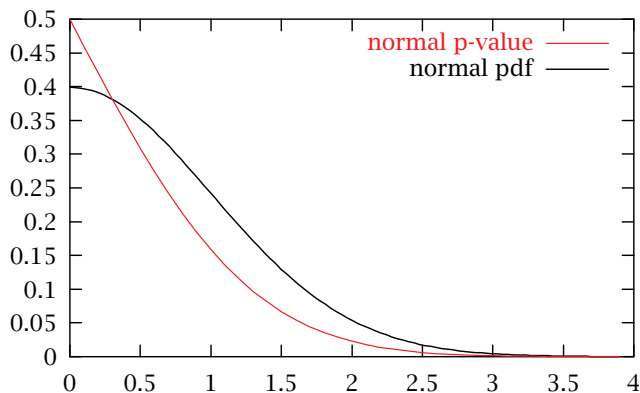
In a two-tailed test at the 5 percent significance level, we fail to reject H_0 if and only if \bar{x} falls within the central 95 percent of the sampling distribution, according to H_0 .

But since 55 exceeds 50 by more than the “maximum error”, 2.45, we can see that, conversely, the central 95 percent of a sampling distribution centered on 55 will not include 50, so a finding of $\bar{x} = 50$ must lead to rejection of the null.

“Significance level” and “confidence level” are complementary. \square

33

Digression



The further we are from the center of the sampling distribution, according to H_0 , the smaller the p-value.

Back to the [main discussion](#).

34