Unknown Proportions and Means

Decisions with Data | Inference for proportions

STAT5002

The University of Sydney

Apr 2025

THE UNIVERSITY OF SYDNEY

CRICOS 00026A TEQSA PRV12057

Decisions with Data

Topics 8 and 9: Confidence intervals and the z-test

Topic 10: The t-test

Topic 11: The two-sample test

Topic 12: χ^2-test

Hypothesis Test: for 0-1 box

Special value of p

- Suppose we have data modelled as a random sample from a 0-1 box, with unknown proportion p of 1s.
- We have seen how to produce a confidence interval, which is an interval containing values of p that the data is consistent with.
- In some scentific scenarios, there is a special value of the parameter (proportion) which might be of interest.
- Example: A company claims that 60% of customers prefer their product ($p = 0.6$). We collect data to assess this claim.
- Instead of estimating p, we may want to test if the data supports or contradicts this special value.

Example

- Suppose that historical data indicates that the proportion of rainy days at Canterbury in March is 0.2.
- Is the March 2024 data (i.e. $\bar{x} = \frac{s}{n} = \frac{13}{31} \approx 0.42$) consistent with this?
    - Let us interpret "consistent" to mean "in the sense of 95% prediction".
- **Solution**: The answer is NO. We can see this two ways

1. The **quick** way:

- We have already computed a 95% confidence interval based on this data: $(0.26, 0.59)$.
- These are the values of the p parameter that the data are consistent with.
    - Since 0.2 is not included, the data is not consistent with 0.2, in this sense.

2. The **slow** way: We can explicitly construct a 95% prediction interval for $\bar{X}$ when the true $p = 0.2$. This would look like:

```r
round(0.2 + c(-1, 1) * 1.96 * sqrt(0.2 * 0.8/31), 2)
```

```
[1] 0.06 0.34
```

- Since this does not include 0.42, we conclude that an observation of 0.42 is *not* consistent with $p = 0.2$ (in this sense).

We will focus on the slow way to introduce hypothesis tests.

False alarm rate / level of significance

- In fact, before we see the data, we can indicate what our conclusion would be for any potential observation:
    - If the observed $\bar{x}$ is within the range (0.06, 0.34) we would conclude "**data is consistent with the hypothesis $p = 0.2$**";
    - If the observed $\bar{x}$ is *outside* the range (0.06, 0.34) we would **"reject" the hypothesis $p = 0.2$**
- This "rejection" statement entails some risk: there is a chance we can reject the hypothesis incorrectly.
    - When the hypothesis is true (i.e true $p = 0.2$) there is a 5% chance $\bar{X}$ lands outside (0.06, 0.34).
    - This is the **false alarm rate** or **level of significance** of our procedure
        - the chance we reject the hypothesis when it is true.
- The smaller the false alarm rate, the more "cautious" we are:
    - We then only reject the hypothesis if there is overwhelming evidence in the data.

Measuring strength of “evidence against”

- What if we instead start with a 99% prediction interval for $p = 0.2$?
- As we have seen, this would give $(0.015, 0.385)$:

```r
round(0.2 + c(-1, 1) * 2.576 * sqrt(0.2 * 0.8/31), 3)
```

```
[1] 0.015 0.385
```

- Our observed $\bar{x} \approx 0.42$ is also outside this, so we would *also* reject the hypothesis $p = 0.2$ at the 1% level of significance.
- How small do we have to make the false alarm rate before we do not reject?

- A 99.9% prediction interval would use a multiplier which has only 0.05% in the upper tail of the standard normal curve:

```r
qnorm(0.9995)
```

```
[1] 3.290527
```

- The corresponding prediction interval is

```r
0.2 + c(-1, 1) * 3.29 * sqrt(0.2 * 0.8/31)
```

```
[1] -0.03636058  0.43636058
```

- **Finally**, this includes the observed value 0.42.
- So we would *not* reject the hypothesis $p = 0.2$ if we used the **super-cautious** 0.1% false alarm rate.

Observed level of significance / P-value.

- We can work out the **exact** false alarm rate at which the observed $\bar{x} = 0.42$ is *right on the edge*.
- This is the level which uses as multiplier the value z such that

$$0.2 + z\sqrt{\frac{0.2 \times 0.8}{31}} = 0.42\,, \quad \text{that is} \quad z = \frac{0.42 - 0.2}{\sqrt{0.2 \times 0.8/31}} \approx 3.06\,.$$

- The desired false alarm rate is simply **twice** the upper tail area beyond 3.06:

```r
2 * pnorm(3.06, lower.tail = F)
```

```
[1] 0.00221337
```

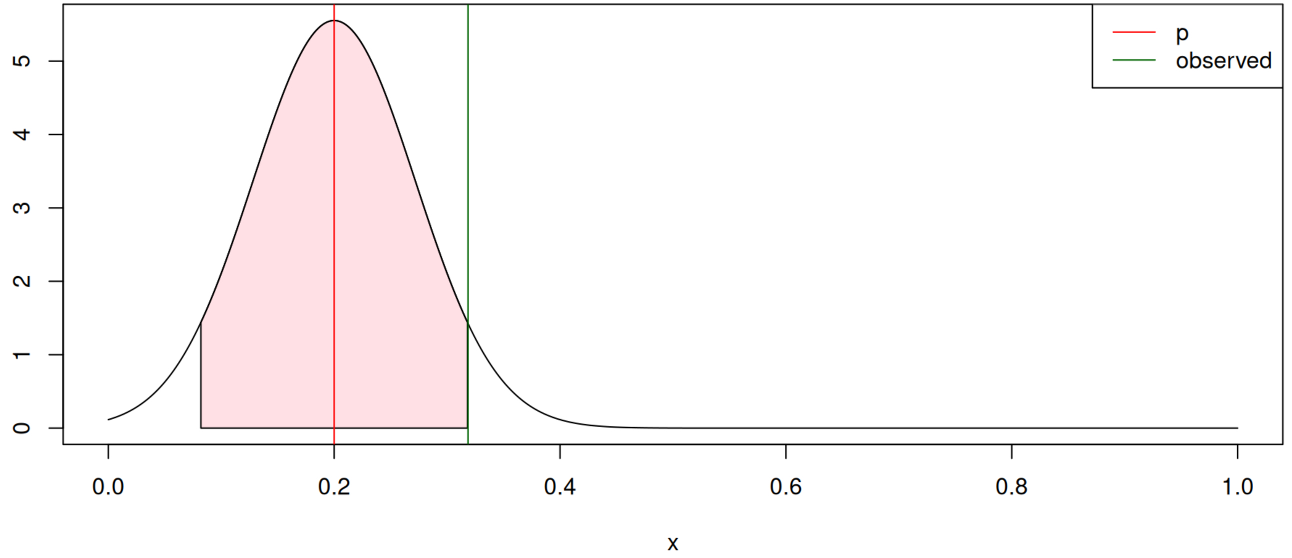
This (small) quantity is the **observed level of significance** or **P-value** based on the data.

Interpreting the P-value

- The smaller the P-value, the stronger the evidence in the data against the hypothesis.
- The P-value may be interpreted as a probability:

The probability of getting as much evidence against the hypothesis as was observed, when the hypothesis is true.

➡ suppose the hypothesis is true, it is the probability of observing something more extreme than the observed sample.



Z-statistic

- Using the normal approximation of the box model, we know that the observed sample mean follows a normal curve.
- The "crucial value of the multiplier"

$$z = \frac{0.42 - 0.2}{\sqrt{0.2 \times 0.8/31}} = \frac{\bar{x} - E_0(\bar{X})}{SE_0(\bar{X})}$$

  where $p_0 = 0.2$ is the *hypothesised value*.
- This simply measure how many SEs away the observed value $\bar{x}$ is from the expected value, converting the observed $\bar{x}$ into *standard units*, assuming the hypothesis is true.
- The value z is in turn the observed value of the (random) **Z-statistic**:

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})}$$

  which is approximately distributed like a draw from a standard normal box **if the hypothesis is true**.

Summary of Z-test procedure

1. Compute the value

$$z = \frac{\bar{x} - E_0(\bar{X})}{SE_0(\bar{X})}$$

   where $E_0(\cdot)$ and $SE_0(\cdot)$ are computed assuming the hypothesis is true.

2. Compute the P-value `2*pnorm(abs(`z`),lower.tail=F)`.
3. Conclude that
   - the data is **consistent with the hypothesised value** at any significance level smaller than the P-value
   - the data is **significantly different from the hypothesised value** at any significance level larger than the P-value

A review of inference for unknown proportions

Hypothesis test of $H_0\colon p = p_0$

- In many contexts a single value p_0 may be of particular interest.
- In such cases, we may formally “test” the hypothesis $H_0\colon p = p_0$ (that the unknown proportion p is equal to the special value p_0).
- The hypothesis $H_0\colon p = p_0$ we test against is called the **null hypothesis**.

Rejecting H_0

- We thus **reject** the null hypothesis $H_0\colon p = p_0$ at the 5% **level of significance** (false alarm rate) if (and only if) the observed sample proportion $\bar{x}$ is **NOT** in the 95% prediction interval for p_0, that is if
  - ➟ $\bar{x} < p_0 - 1.96\sqrt{\frac{p_0(1-p_0)}{n}}$ *or*
  - ➟ $\bar{x} > p_0 + 1.96\sqrt{\frac{p_0(1-p_0)}{n}}$; equivalently if

$$|z| = \frac{|\bar{x} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} > 1.96\,.$$

  - ➟ In such a case we also say the observed $\bar{x}$ is **significantly different to p_0** (at the 5% level of significance).

Not rejecting H_0, but not accepting it either

- If $\bar{x}$ lands within the prediction interval, i.e. if

$$|z| = \frac{|\bar{x} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq 1.96\,,$$

  we say the data is **consistent with H_0** (at the 5% level of significance).

- We do not "accept" $H_0\colon p = p_0$, since a single observation does not "prove a hypothesis true":
  - ➟ It is consistent with a whole set of values for p: i.e. all values in the 95% confidence interval!

Multiplier / Critical value

- Prediction intervals have a multiplier that is used in building confidence intervals.
- The confidence level depends on the "multiplier" of the standard error:
  - ➟ 95% confidence uses the multiplier 1.96.
- The significance level is also controlled by this same multiplier: we reject if the observed value of the **Z-statistic**

$$Z = \frac{\bar{X} - E_0\left(\bar{X}\right)}{SE_0(\bar{X})} = \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

exceeds (in **absolute value**) the multiplier (also called "critical value"):

- ➟ the 5% level of significance also uses the critical value 1.96.

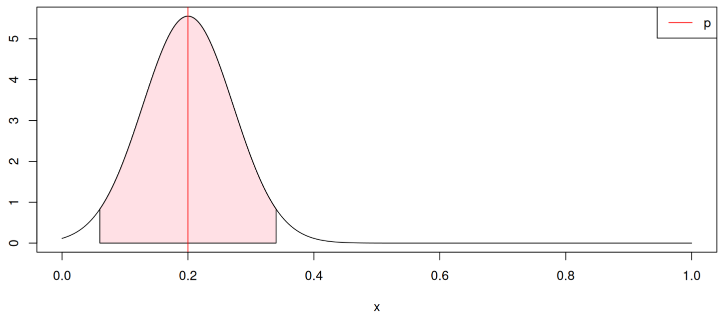
Different confidence/significance levels

- The multiplier/critical value cuts off a certain area in the upper tail under the standard normal curve.
- In general, suppose we have a significance level $0 \leq \alpha \leq 1$.
  - Then the multiplier/critical value cuts of an area of $\alpha/2$ in the upper tail of the standard normal curve.
- Multipliers/critical values for common confidence/significance levels are:

Sig. level	Conf. level	Upper tail area	Multiplier/critical value	R command
0.05	95%	0.025	1.960	`qnorm(0.975)`
0.02	98%	0.010	2.326	`qnorm(0.99)`
0.01	99%	0.005	2.576	`qnorm(0.995)`
0.001	99.9%	0.0005	3.291	`qnorm(0.9995)`
0.0001	99.99%	0.00005	3.891	`qnorm(0.99995)`

Observed significance level / P-value

- Using a single significance level either rejects, or does not reject H_0.
  - ➟ It fails to convey the “strength of evidence” against the hypothesis.
- One can also quote the **observed level of significance** or **P-value** associated with an observed $\bar{x}$.
- **Smaller** P-value means **more evidence against the hypothesis**.
  - ➟ the probability of observing something more extreme than the observed sample (under H_0).



- It is given by `2*pnorm(abs(`z`), lower.tail=F)` and measures the chance a random draw from a standard normal box
  - ➟ *Either* exceeds $|z|$ *or* is less than $-|z|$.
- The “two-sided” nature of this calculation reflects the fact that alternative values of p both *above* and *below* the hypothesised value p_0 are “equally of interest”.

One-sided tests

Which alternatives are of interest?

In many practical hypothesis-testing scenarios, values both above and below the hypothesised value p_0 might be of interest.

Examples:

- $p =$ proportion of days with rain in March: is climate change increasing or decreasing rain in March?
  - ➡ (p_0 represents historical proportion of days with rain in March)
- $p =$ proportion of patients showing improvement using a new drug: is the new drug better or worse than the current standard treatment?
  - ➡ (p_0 represents proportion of patients showing improvement with current standard treatment)

One-sided alternatives: production lines

In some scenarios, only alternatives in one particular direction are of interest.

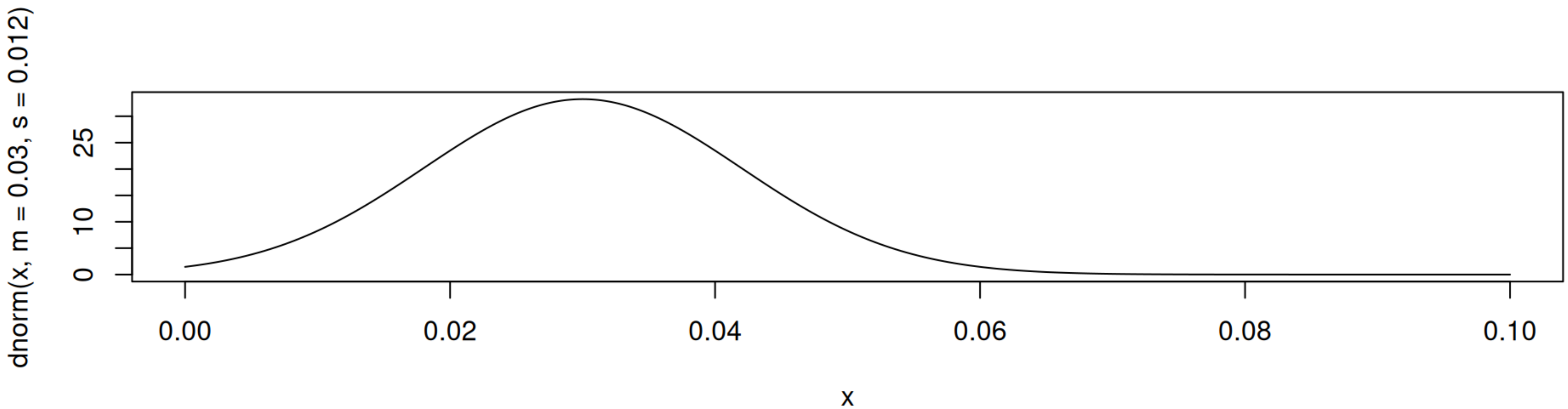
- Suppose a production line produces items at a rate of 5000 per day.
- The process occasionally produces faulty items.
- It is deemed "acceptable" if 3% of the items are faulty.
- As a quality control measure, once a week a random sample of $n = 200$ items is taken and the proportion of faulty items $\bar{x}$ determined.
- If there is evidence that the "failure rate" is higher than 3%, they stop the production and repair the machines.
  - ➠ This is a costly process, so it should only be done if the evidence is "clear".
- How should such a test be performed so that the false alarm rate (the chance of needless shutdown) is no more than 1%?

Formal procedure: null and alternative hypotheses

- We formally specify the test as follows:
- The parameter p represents the actual proportion of faulty items being produced.
- Our **null hypothesis** is $H_0: p = 0.03$.
  - ➠ This represents "nothing interesting going on".
- We declare our **alternative hypothesis** to be $H_1: p > 0.03$.
  - ➠ These are the alternatives we are "trying to detect".
- The direction of the alternative suggests what procedure to use.

Distribution of $\bar{X}$ when H_0: $p = 0.03$ is true

- The null hypothesis thus represents when the production process is "running normally".
  - ➟ We would expect the (random) proportion of faulty items to be $p_0 = 0.03$.
- In fact we would expect its distribution to be normal shaped, with
  - ➟ $E_0(\bar{X}) = p_0 = 0.03$ and
  - ➟ $SE_0(\bar{X}) = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.03\times 0.97}{200}} \approx 0.012$.



Only reject for very large $\bar{X}$

- We are only interested in alternatives where the expected proportion of defects $p > 0.03$.
- We should only reject H_0 if we get a "larger than expected" proportion of defects.
  - ➠ If the observed proportion of defects is *less* than expected, we just "carry on"!
- We should thus reject if the sample proportion $\bar{X}$ takes a value $\bar{x}$ **larger** than some critical value c.
- How should the critical value c be chosen?
  - ➠ So that the *probability of incorrect rejection* is $\alpha = 0.01$ (the desired false alarm rate)!

Standard normal critical value

- The value that cuts off 1% in the upper tail of the standard normal curve is 2.326:

```r
round(qnorm(0.99), 3)
```

```
[1] 2.326
```

- Thus for any normal-shaped box, the upper 1% of values are those more than 2.326 SDs above the mean.
- Under H_0, the sample proportion is like a draw from a normal box with mean 0.03 and SD 0.012.
- The critical value should thus be

```r
round(0.03 + 2.326 * 0.012, 3)
```

```
[1] 0.058
```

- So if we reject for a sample proportion more than 0.058 (i.e., about 12 or more out of 200), the false alarm rate is 1%.

In terms of the Z-statistic

- The Z-statisic for testing $H_0: p = 0.03$ is then

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})} = \frac{\bar{X} - 0.03}{0.012}.$$

- If we observe Z to take the value

$$z = \frac{\bar{x} - 0.03}{0.012}$$

we reject H_0 (at the 1% level of significance) if (and only if) $z > 2.326$.

P-value

- The observed level of significance will then be given by `pnorm(`z`, lower.tail=F)`.
  - Note this is the same as `pnorm(`$\bar{x}$`, m=0.03, s=0.12, lower.tail=F)`.
- This the chance that a random draw from a standard normal box exceeds the observed value z of the Z-statistic.
  - We do not worry about any lower tail area here.

Specify conclusions before seeing the data

- We can thus indicate what we will conclude *before seeing the data*.
  - ➠ This is important to prevent “data snooping” i.e. letting the data suggest the procedure.
- The output below shows some potential outcomes and corresponding P-values

```
       s  xbar         z   P.val
 [1,]  6 0.030 0.0000000 0.50000
 [2,]  7 0.035 0.4166667 0.33846
 [3,]  8 0.040 0.8333333 0.20233
 [4,]  9 0.045 1.2500000 0.10565
 [5,] 10 0.050 1.6666667 0.04779
 [6,] 11 0.055 2.0833333 0.01861
 [7,] 12 0.060 2.5000000 0.00621
 [8,] 13 0.065 2.9166667 0.00177
 [9,] 14 0.070 3.3333333 0.00043
[10,] 15 0.075 3.7500000 0.00009
[11,] 16 0.080 4.1666667 0.00002
```

Summary

- If we are testing $H_0\colon p = p_0$ based on an observed proportion $\bar{x}$, or equivalently observed Z-statisic

$$z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

  we can identify which procedure we should use by identifying which of these questions is appropriate:
- If we ask "*Is $\bar{x}$ significantly* ***greater*** *than p_0?*", we should use the (one-sided) alternative $H_1\colon p > p_0$ and compute the P-value using `pnorm(`z`,lower.tail=F)`;
- If we ask "*Is $\bar{x}$ significantly* ***less*** *than p_0?*", we should use the (one-sided) alternative $H_1\colon p < p_0$ and compute the P-value using `pnorm(`z`)`;
- If we ask "*Is $\bar{x}$ significantly* ***different*** *to p_0?*", we should use the (two-sided) alternative $H_1\colon p \neq p_0$ and compute the P-value using `2*pnorm(abs(`z`),lower.tail=F)`.