1.5 Measures of Variation

Procedure for obtaining the quartiles:

1. Order the n data values from smallest to largest.
2. The 2nd quartile Q_2 is the median of the whole data set.
3. If n is odd, the 1st quartile Q_1 is the median of the smallest $\frac{n-1}{2}$ observations and the 3rd quartile Q_3 is the median of the largest $\frac{n-1}{2}$ observations.
4. If n is even, Q_1 is the median of the smallest $\frac{n}{2}$ observations, and Q_3 is the median of the largest $\frac{n}{2}$ observations.

1.5 Measures of Variation

Taking the square root, we have $\sigma = \sqrt{\frac{1}{n}\sum_i d_i^2}$

This is called the population standard deviation, and it is denoted by σ.

However, it is a common practice nowadays to modify the formula and replace the denominator n with $n-1$,

i.e. $s = \sqrt{\frac{1}{n-1}\sum_i d_i^2} = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$

This is called the sample standard deviation, and it is denoted by s.

2. An alternative version of the formula is

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

2.3 Conditional Probability

Recall that : $P(C \text{ and } D) = 0.11$

$$\therefore P(C|D) = \frac{P(C \text{ and } D)}{P(D)} = \frac{0.11}{0.47} = 0.23$$

In general, for events X and Y the conditional probability of X given that Y has occurred is

$$P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)}$$

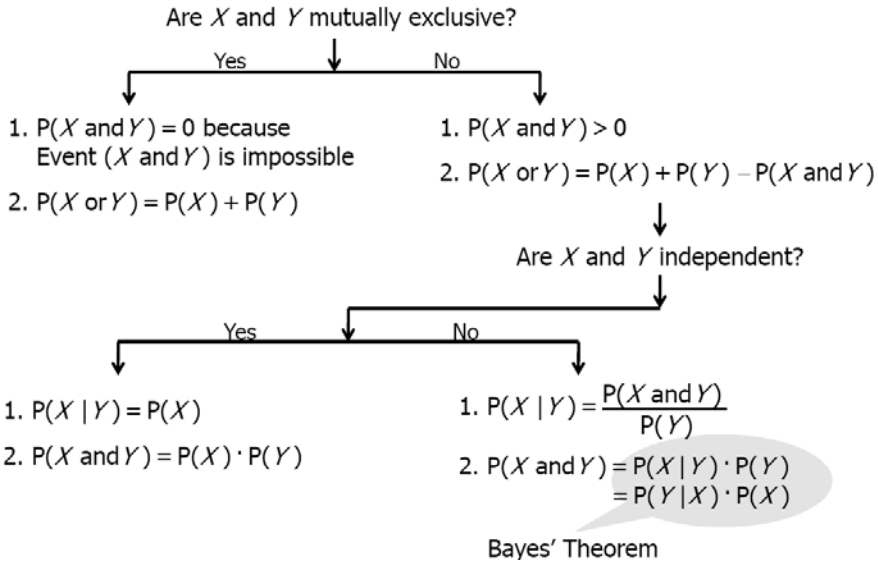
2.4 Bayes' Theorem

In the last example, in which the order of the condition is reversed (i.e. obtaining P(M|E) from P(E|M)), illustrates Bayes' Theorem.

In general, for any 2 events X and Y

P(X and Y) = P(Y and X) $\Rightarrow$ $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$

Summary For Probability Calculations



3.1 Random Variables

Variance of a r.v.

Recall that variance is a measure of spread. For a sample of n observations the sample variance is defined as

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

The variance of a r.v. X is defined as

$$\begin{aligned} \text{var}(X) &= p_1(x_1-\mu)^2 + p_2(x_2-\mu)^2 + \cdots + p_n(x_n-\mu)^2 \\ &= \sum_{i=1}^{n} p_i(x_i-\mu)^2 = E\left[(x-\mu)^2\right] \end{aligned}$$

where p_i = probability that x_i occurs.

It represents the theoretical limit of the sample variance s^2 as the sample size n becomes very large. Var(X) is often denoted by σ^2.

A simpler formula for var(X) is as follows:

$$\text{var}(X) = (p_1 x_1^2 + \ldots + p_n x_n^2) - \mu^2$$

$$= E[X^2] - \mu^2 \quad \text{or} \quad E[X^2] - (E[X])^2$$

Empirical quantity (data based)	Theoretical quantity (mathematical)	Remarks
(a) Relative freq. of x_i is $\frac{f_i}{n}$	$P[X = x_i] = p_i$	$\frac{f_i}{n} \to p_i$ as $n \to \infty$
(b) $\sum_{i=1}^{n} \frac{f_i}{n} = 1$	$\sum_{i=1}^{n} p_i = 1$	
(c) Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i f_i$	Expectation: $E(X) = \sum_{i=1}^{n} p_i x_i = \mu$	$\bar{X} \to E(X)$ as $n \to \infty$
(d) Variance: $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2 f_i}{n-1}$	$\text{var}(X) = \sum_{i=1}^{n} (x_i - \bar{X})^2 p_i = \sigma^2$	$s^2 \to \text{var}(X)$ as $n \to \infty$

Expectation $E(S^2_{n-1}) = \sigma^2$

Proof: The unbiased estimate of the variance is

$$s^2 = \frac{1}{n-1}[\sum X^2 - \frac{(\sum X)^2}{n}]$$

Recall that $\text{var}(X) = E(X^2) - [E(X)]^2$

So $E(X^2) = \text{var}(X) + [E(X)]^2$ —— (1)

$= \sigma^2 + \mu^2$ —— (2)

Case *n*:

If $s_n^2 = \frac{1}{n}[\sum X^2 - \frac{(\sum X)^2}{n}]$

We have $E(s_n^2) = E\left(\frac{1}{n}[\sum X^2 - \frac{(\sum X)^2}{n}]\right)$

$= \frac{1}{n}\left(\sum E(X^2) - \frac{1}{n}E[(\sum X)^2]\right)$

3.2 Prob. Distribution of Discrete r.v.

Let *X* be the r.v. equal to the total number of successes in *n* trials. To calculate the probability of obtaining *x* successes, it can be shown that

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

- $\binom{n}{x}$: # arrangements of *x S*'s and (*n*-*x*) *F*'s
- p^x: prob. of *x S*'s
- $(1-p)^{n-x}$: prob. of (*n*-*x*) *F*'s

The distribution of the count of successes is called the binomial distribution with two parameters, *n* and *p*. We say $X \sim B(n, p)$.

X	0	1	...	n
$P(X=x)$	${}^nC_0 p^0 (1-p)^n$	${}^nC_1 p^1 (1-p)^{n-1}$	...	${}^nC_n p^n (1-p)^{n-n}$

3.3 Prob. Distribution of Continuous r.v.

The expected value of a continuous r.v. is defined as:

$$E[X] = \int_{x=-\infty}^{x=\infty} x\, f(x)\, dx = \mu$$

and its variance is:

$$\text{var}[X] = \int_{x=-\infty}^{x=\infty} (x-\mu)^2 f(x)\, dx = E[X^2] - E[X]^2$$

where $E[X^2] = \int_{x=-\infty}^{x=\infty} x^2 f(x)\, dx$

3.3 Prob. Distribution of Continuous r.v.

Uniform Distribution

One example of a pdf for a continuous r.v. is the uniform continuous distribution.

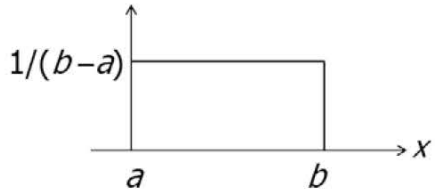
X can take any real value between *a* and *b* with uniform probability over this interval.

Notation: $X \sim U(a,b)$.

Total area = 1

Length = $b - a$

Height = $1/(b-a)$

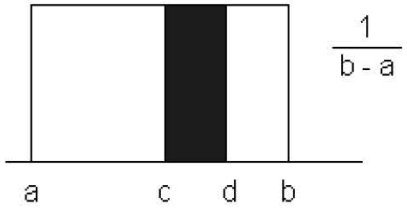


3.3 Prob. Distribution of Continuous r.v.

Thus the pdf is $f(X)=\frac{1}{b-a}$ for $a \le X \le b$

For any values c and d between a and b

$$P(c < X < d) = \frac{1}{b-a}(d-c)$$



Exercise: Obtain the CDF.

3.3 Prob. Distribution of Continuous r.v.

Normal Distribution

Histograms, as the no. of observations increases, can be approximated by a continuous function $f(X)$.

A common function is the normal (bell-shaped) curve. Its pdf depends on μ and σ, and is given by:

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \text{ for } -\infty < X < \infty$$

The normal distribution is by convention written as $X \sim N(\mu, \sigma^2)$. Eg: if $X \sim N(10, 16)$ then this implies that $\mu=10$ and $\sigma^2=16$, and (hence $\sigma=4$).

3.4 Approx. to Binomial Distribution

If $X \sim B(n, p)$ where n is large and p not too near 0 or 1, then X can be approximated by a normal distribution with $E(X)=np$ and $\text{var}(X)=npq$, where $q=1-p$.

So $Z=\frac{X-np}{\sqrt{npq}}$ is approximately $N(0, 1)$.

This approximation is reasonably good when $np > 5$ and $n(1-p) > 5$.

4.2 The Sampling Distribution

Properties of the Sampling Distribution

1. The sampling distributions of $\bar{X}$ have the same expected value μ, regardless of sample size n.
2. The variance of the sampling distribution is:

$$\text{var}(\bar{X})=\text{var}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)=\frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} x_i\right)$$

$$=\frac{1}{n^2}\text{var}(x_1+\ldots+x_n) \quad =\frac{1}{n^2}\left(\sigma^2+\ldots+\sigma^2\right)=\frac{1}{n^2}n\sigma^2$$

$$=\frac{\sigma^2}{n}$$

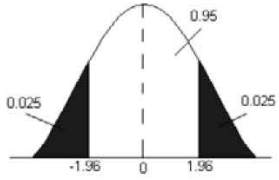
3. The square root of this variance, $\sqrt{\text{var}(\bar{X})}$, is called the "standard error" of $\bar{X}$: SE $= \frac{\sigma}{\sqrt{n}}$
4. The variance is inversely proportional to n. As n increases, the distribution of $\bar{X}$ becomes narrower, i.e., the $\bar{X}$'s cluster more tightly around μ.
5. The sampling distribution of $\bar{X}$ is approximately $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ even if the distribution of the X values were not normal. This followed from the Central Limit Theorem.

4.2 The Sampling Distribution

To obtain the 95% confidence limits, from the tables for $N(0, 1)$, we get: $P(Z<1.96)=0.975$

Hence $P(-1.96<Z<1.96)=0.95$

i.e. $$P\left(-1.96 < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$



This can be rearranged to obtain 95% confidence limits for μ.

$$\left.\begin{aligned} L &= \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \\ U &= \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \end{aligned}\right\} \text{95\% confidence limits}$$

4.3 Hypothesis Testing

The Confidence Interval Approach

This approach is to find a confidence interval for μ.

A 95% confidence interval is given by

$14.85 \pm 1.96 \times \frac{0.3}{\sqrt{50}}$ i.e. (14.77, 14.93) cm

Interpretation: with 95% confidence the interval (14.77, 14.93) contains the population mean μ of all bolts produced by the process.

As the interval does not contain 15.0, the data are not consistent with the hypothesis that μ =15. That is, the data do not support the hypothesis that the average length of all bolts is 15 cm.

Confidence intervals with other probabilities, e.g. 0.9 or 0.99, can be calculated similarly. For the above example, a 99% confidence interval will be

4.3 Hypothesis Testing

Steps for Hypothesis Testing using the *p*-value Approach

In general the steps for the *p*-value approach is as follows:

1. Formulate the null hypothesis and an appropriate alternative hypothesis, eg: $H_0: \mu \geq 15$, $H_A: \mu < 15$
2. Specify the level of significance α
3. Specify the test statistic (eg: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$)
4. Calculate the value of the test statistic and the corresponding *p*-value from the data (e.g. if z = 2.05, then *p*-value = 0.0202 for a one-tailed test)
5. Compare the *p*-value obtained in step 4 with the level of significance α in step 2. If *p*-value $\leq \alpha$, reject the null hypothesis, otherwise accept the null hypothesis. (e.g. if *p*-value = 0.0202 ≤ 0.05, the null hypothesis must be rejected).

4.4 Test Concerning Means

Eg:

The manager of a trucking firm suspects the claim of a tyre salesman that the average lifetime of his tyres is at least 28,000 miles. To check the claim, the firm mounts 40 of these tyres on its trucks and gets a mean lifetime of 27,563 miles with a standard deviation of 1,348 miles. What can the manager conclude if the level of significance is 0.01?

4.4 Test Concerning Means

Solution

1. $H_0: \mu \geq 28{,}000$, $H_A: \mu < 28{,}000$ (one-tailed test)
2. $\alpha = 0.01$
3. The test statistic is $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

   Since σ is not known and n>30, we can approx it with s.
4. Substitute $\mu_0 = 28{,}000,\ \bar{x} = 27{,}563,\ n = 40,\ s = 1{,}348$

   into the formula for z, we get $z = \dfrac{27{,}563 - 28{,}000}{1{,}348/\sqrt{40}} \approx -2.05$

   and from the Z-table, the *p*-value, the area under the curve to the **left** of z = −2.05, is 0.0202.
5. Since 0.0202 > α, the null hypothesis cannot be rejected. In other words, the manager's suspicion that μ < 28,000 miles cannot be confirmed by the data.

4.5 Test Concerning Means (Small Samples)

Eg:

Bottles of soft drinks are meant to contain 300ml. A sample of n = 10 bottles were measured and the contents were: 299, 276, 283, 301, 297, 281, 300, 291, 295, & 291.

(a) Test the null hypothesis $H_0:\mu = 300$ against $H_A:\mu \neq 300$.

(b) Find a 90% confidence interval for μ.

(a) 1. $H_0:\mu$ = 300, $H_A:\mu \neq 300$ (two-tailed test)

2. α = 0.05 (by default)

3. The test statistic is $t_r = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

4. Subst n=10, $\bar{x}$ =291.4, μ_0=300, s=8.72, r=9

$$t_9 = \frac{291.4 - 300}{8.72/\sqrt{10}} = 3.12$$

p-value = $P(t_9<-3.12$ or $t_9>3.12) = 2 \times P(t_9>3.12)$

and from the t-table,

$$2 \times P(t_9 > 2.821) > 2 \times P(t_9 > 3.12) > 2 \times P(t_9 > 3.25)$$

hence $0.02 > 2 \times P(t_9 > 3.12) > 0.01$

$$0.02 > p - \text{value} > 0.01$$

5. Since p-value < 0.02 which is < 0.05, the null hypothesis must be rejected. The data provide evidence against $H_0:\mu$ = 300.

   Note that the t-table for r = 9 does not have a probability for the value 3.12. The tabulated value to choose is 2.821 and 3.25.

(b) To find a 90% C.I. for μ, we first need to find the value t such that $P(-t<t_9<t) = 0.9$

To do this, an area of 0.05 must be left in each tail of the t_9-distribution. From the t-table, t must take the value 1.833.

i.e $P(-1.833 < t_9 < 1.833) = 0.9$

Hence the 90% C.I. is given by:

$$\bar{X} \pm 1.833\frac{s}{\sqrt{n}} = 291.4 \pm 1.833 \times \frac{8.72}{\sqrt{10}} = \text{(286.3, 296.5)}$$

i.e. with 90% confidence the mean contents of the bottles is in the interval (286.3, 296.5) ml. As this interval does not contain the value 300, the data do not support the hypothesis that the true mean content of all drink bottles is 300ml.

Note that for a 95% C.I. the calculation would be $\bar{X} \pm 2.262\frac{s}{\sqrt{n}}$ instead.

4.6 Test Concerning Proportions

It follows that:

expected value, $E(\hat{p}) = \frac{1}{n}E(X) = \frac{np}{n} = p$

variance, $\text{var}(\hat{p}) = \frac{1}{n^2}\text{var}(X) = \frac{npq}{n^2} = \frac{pq}{n}$

standard deviation, $SD(\hat{p}) = \sqrt{\text{var}(\hat{p})} = \sqrt{\frac{pq}{n}}$

If $np > 5$ and $n(1-p) > 5$, $X \sim B(n, p)$ can be approximated by the Normal distribution $\hat{p} \sim N(p, \frac{pq}{n})$

This gives us the test statistic $Z = \frac{\hat{p} - p}{SD(\hat{p})}$ for making inferences about *p*.

4.6 Test Concerning Proportions

Eg: A company claims to have 40% of the market for some product. A survey shows 38 out of 112 buyers (i.e. 34%) purchased this brand. Are these data consistent with the company's claim or the survey result of 34% significantly different to the company's claim of 40%?

1. $H_0 : p = 0.4$, $H_A : p \neq 0.4$
2. α = 0.05
3. The test statistic is $z = \frac{\hat{p} - p}{SD(\hat{p})}$

4.6 Test Concerning Proportions

4. Given n=112, p=0.4, q=0.6, calculate:

$$\hat{p} = \frac{38}{112}, \quad SD(\hat{p}) = \sqrt{\frac{pq}{n}}, \quad z = \frac{\hat{p} - p}{SD(\hat{p})} \approx -1.31$$

and from the Z-table we find:

p-value = P(Z<-1.31 or Z>1.31)

= 2xP(Z>1.31) = 2x0.0951 ≈ 0.19

5. Since 0.19 > α, the null hypothesis cannot be rejected. In other words, the data are consistent with the claim that the true population proportion of buyers purchasing this brand is 40%.

4.7 Test Concerning Proportions (Small Samples)

Eg: Suppose only 3 of the first 21 babies were boys. Would that have suggested that the sex ratio for IVF was different from 1:1?

Solution

1. H_0: p=0.5, H_A: p≠0.5
2. α = 0.05
3. The test statistic is the number of boys: X = 3
4. For n=21, p=0.5, we have $E(X)=np$=10.5.

$\therefore$ p-value = 2xP($X \leq 3$)

$$= 2 \times \left[\binom{21}{0} \tfrac{1}{2}^{0} \tfrac{1}{2}^{21} + \binom{21}{1} \tfrac{1}{2}^{1} \tfrac{1}{2}^{20} + \cdots + \binom{21}{3} \tfrac{1}{2}^{3} \tfrac{1}{2}^{18} \right] \approx 0.0015$$

5. Since the p-value is less than $\alpha = 0.05$, the null hypothesis must be rejected. We conclude that the sex ratio differed from 1:1.

Note that the observed proportion of boys, 1/7, was the same for both examples. The p-value was affected by the total number, n, of babies or equivalently, by the expected numbers of boys and girls, np and $n(1 - p)$.