Estimation, Confidence Intervals and Tests Using a Normal Distribution Cheat Sheet

Statistic, estimator, and bias

For a large population in which the population parameters, mean and standard deviation are unknown, it is possible to estimate them from the parameters of a random sample.

Let's say that X, the weight of a batch of apples is a random variable with unknown mean (μ) and standard deviation (σ). A random sample of 10 apples are taken.
$X_1, X_2, X_3, X_4 \ldots X_{10}$ are observations of the random sample. They are independent random variables and share the same distribution as X. The mean and standard deviation of the random sample can be calculated.

A **statistic**, T, is any function of the random variable X which does not contain any unknown parameter. Statistics derived from a sample can be used as **estimators** for population parameters. T is also a random variable and since it can be different for every random sample taken, the sampling distribution of T is also its probability distribution.

When a statistic T is used to estimate a population parameter θ and $\mathrm{E}(T) = \theta$, then T is said to be an **unbiased estimator**. However, more often $\mathrm{E}(T) \neq \theta$ and this is known as a **biased estimator**. The difference between the estimate and the population parameter is known as the **bias**.

Example 1: The table shows the lowest temperature of the day in °C, x, on 25 random days.

x	3	4	5	6	7	8
Number of days	5	4	3	5	4	4

(a) Calculate the unbiased estimates of mean and variance of x.

Find $\sum x$ and $\sum x^2$.	$\sum x = 136$ $\sum x^2 = 816$
Calculate the mean.	$\hat{u} = \bar{x} = \frac{136}{25} = 5.44$
Calculate the variance.	$\sigma^2 = s_x^2 = \frac{816 - 25(5.44)^2}{24} = \frac{238}{75} = 3.17$

(b) Another 20 days were sampled with a mean of 6.22 and a variance of 4.07. Combine these two samples and calculate the new unbiased estimates of mean and variance.

Find $\sum y$ and $\sum y^2$ of the new sample from the given mean and variance.	$\sum y = 6.22 \times 20 = 124.4$ $\sum y^2 = 4.07 \times 19 + 20(6.22)^2 = 851.098$
Find $\sum w$ and $\sum w^2$ for the new combined sample.	$\sum w = \sum x + \sum y = 136 + 124.4 = 260.4$ $\sum w^2 = \sum x^2 + \sum y^2 = 816 + 851.098 = 1667.098$
Find the new estimate of mean.	$\hat{u} = \bar{w} = \frac{260.4}{45} = 5.79$
Find the new estimate of unbiased variance.	$\sigma^2 = s_w^2 = \frac{1667.098 - 45(5.79)^2}{44} = 3.60$

Standard error of the mean

As sample size, n, increases, the variance of the estimator of the population mean, μ, decreases. This is also known as the standard error of the mean and it tells us how useful the estimator is.

The standard error of mean is $\frac{\sigma}{\sqrt{n}}$.

Example 1 (continued):

(c) Calculate the standard error of mean.

The most accurate estimate, which is the one from the largest sample size, should be used.	$s_w^2 = 3.60$
Substitute into the formula: standard error of mean: $\frac{\sigma}{\sqrt{n}}$.	$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{3.60}}{\sqrt{45}} = 0.283$

Confidence Intervals

A confidence interval is the range of values that is likely to contain the real population parameter at the probability stated. You can calculate the confidence interval using the formula:

$$\mathrm{CI} = \bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

The value of z is derived from the value of the tables from standard normal distribution. For example, the corresponding value for 95% confidence intervals is 1.96.

The width of the confidence interval can be calculated by:

$$2 \times z \times \frac{\sigma}{\sqrt{n}}$$

Example 2: In a population with a normal distribution with variance 16 , a random sample of size 9 was taken. The mean of the sample is 140.

(a) Calculate the 95% confidence interval for the population mean μ.

Use the normal distribution table to find value of z for 95% confidence interval.	1.96
Substitute into the formula to find upper limit: Upper limit $= \bar{x} + z \times \frac{\sigma}{\sqrt{n}}$	Upper limit $= 140 + 1.96 \times \frac{4}{\sqrt{9}}$ $= 142.6$
Substitute into the formula to find lower limit: Lower limit $= \bar{x} - z \times \frac{\sigma}{\sqrt{n}}$	Lower limit $= 140 - 1.96 \times \frac{4}{\sqrt{9}}$ $= 137.4$
Write the confidence interval for the population mean in brackets	$(137.4, 142.6)$

(b) Calculate the width of the 99% confidence interval.

Use the normal distribution table to find value of z for 99% confidence interval	2.5758
Substitute into the formula: $2 \times z \times \frac{\sigma}{\sqrt{n}}$	$2 \times 2.5758 \times \frac{4}{\sqrt{9}} = 6.8688$

(c) Calculate the smallest sample size needed for a 99% confidence interval with width less than 5.

Width of 99% confidence interval:	$5 > 2 \times 2.5758 \times \frac{4}{\sqrt{n}}$ $5 > \frac{20.6064}{\sqrt{n}}$ $\sqrt{n} > \frac{20.6064}{5}$ $n > 16.98$
Round up to the nearest whole number.	$n = 17$

Central Limit Theorem

The central limit theorem states that regardless of the population's distribution, as the sample size increases, the distribution of the mean approximates a normal distribution.

In other words, if $X_1, X_2, X_3 \ldots X_n$ is a random sample of size n taken from a large population with mean μ and variance σ^2, then $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$.

Hypothesis Testing

Hypothesis testing can be used to test for the mean of a normal distribution.

The test statistic for population mean is:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where μ is the value given by the null hypothesis.

Normal distribution is then used to see if the test statistic falls in the critical region.

Example 3: A factory produces an energy drink which volume is distributed normally with a standard deviation of 5mL. They claim that the mean volume of the drink in each bottle is 250mL. Jason suspects that volume of the drink in the bottle than what the factory claims. He buys a random sample of 16 bottles of the drink and the mean volume in each bottle is 248.2mL.

(a) Test whether Jason's claim is supported at 5% significance level.

State null and alternative hypotheses. (Note: this is a one-tailed test because it is thought that the actual volume is LESS THAN what the factory claimed.)	$\mathrm{H}_0: \mu = 250$ $\mathrm{H}_1: \mu < 250$
Find the test statistic using the formula $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$Z = \frac{248.2 - 250}{\frac{5}{\sqrt{16}}}$ $= -1.44$
Check critical region at 5% S.L.	$Z < -1.96$
Conclusion.	Test statistic is not in the critical region so null hypothesis is accepted. There is insufficient evidence to support Jason's claim.

(b) State the importance of the central limit theorem in this test.

The central limit theorem is used to assume that the mean of the volume of the sampled drinks are normally distributed.