6.4 Distributions based on normal random samples

In this section, we develop the sampling distribution of the sample variance S^2, the joint distribution of $\bar{X}$ and S^2, and the distributions of other important statistics when sampling from a normal distribution.

The joint sampling distribution of $\bar{X}$ and S^2

For a random sample $X_1, \ldots, X_n$, the sample variance S^2 is defined as a rv by

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

This can be used to calculate an estimate of σ^2 when the population μ is unknown. S^2, like any statistic, varies in value from sample to sample. To establish the sampling dist. of S^2 when sampling from a normal population, we first need the following result.

Theorem. If $X_1, \ldots, X_n$ form a random sample from a normal distribution, then $\bar{X}$ and S^2 are independent.

Proof. Consider the covariances between the sample mean and the deviations from the sample mean. Using the linearity of the covariance operation,

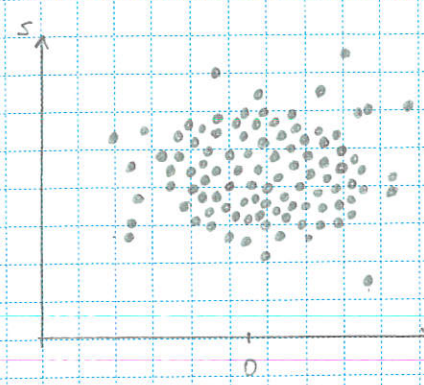
$$\begin{aligned} \mathrm{Cov}(X_i - \bar{X}, \bar{X}) &= \mathrm{Cov}(X_i, \bar{X}) - \mathrm{Cov}(\bar{X}, \bar{X}) \\ &= \mathrm{Cov}\left(X_i, \frac{1}{n}\sum_{k=1}^{n} X_k\right) - V(\bar{X}) \\ &= \frac{1}{n}\mathrm{Cov}(X_i, X_i) + \frac{1}{n}\sum_{k \neq i} \mathrm{Cov}(X_i, X_k) - V(\bar{X}) \end{aligned}$$

Since X_i's are independent, $\mathrm{Cov}(X_i, X_k) = 0$ for $k \neq i$.

$$\begin{aligned} &= \frac{1}{n} V(X_i) + 0 - V(\bar{X}) \\ &= \frac{1}{n}\sigma^2 - \frac{\sigma^2}{n} \\ &= 0 \end{aligned}$$

This shows that $\bar{X}$ is uncorrelated with all the deviations of the observations from their mean. Both $\bar{X}$ and $X_i - \bar{X}$ are linear combinations of the independent normal observations, so their joint dist. is bivariate normal. Then, $\bar{X}$ and S^2 are independent as S^2 is composed of $X_i - \bar{X}$. ∎

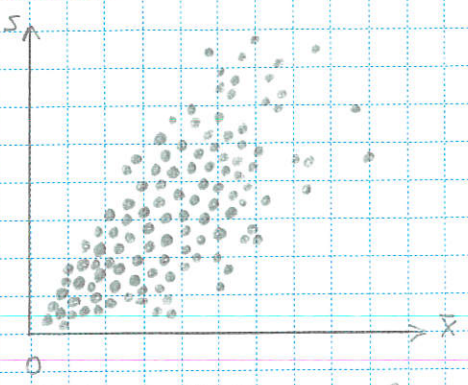
Example. Consider selecting 1000 samples of size $n=5$ from a particular population dist., calculating $\bar{x}$ and s for each sample, and then plotting the resulting $(\bar{x}, s)$ pairs.



Plot of $(\bar{x}, s)$ pairs for samples from a normal distribution

↓

The elliptical pattern shows that no relationship between $\bar{x}$ and s. That is, $\bar{X}$ and S (equivalently $\bar{X}$ and S^2) are independent.



Plot of $(\bar{x}, s)$ pairs for samples from an exponential dist. with mean 1. This plot shows a strong relationship between $\bar{X}$ and S.

Now, let us derive the sampling distribution of S^2 when sampling from a normally distributed population.

Proposition. If $X_1, \dots, X_n$ form a random sample from a $N(\mu, \sigma)$ distribution, then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Proof. Consider the summation $\sum (X_i - \mu)^2$. We can write

$$\sum (X_i - \mu)^2 = \sum [(X_i - \bar{X}) + (\bar{X} - \mu)]^2$$

$$= \sum (X_i - \bar{X})^2 + \underbrace{2(\bar{X} - \mu)\sum (X_i - \bar{X})}_{=0} + \sum (\bar{X} - \mu)^2$$

dividing by σ^2.

$$\Rightarrow \underbrace{\sum \left(\frac{X_i - \mu}{\sigma}\right)^2}_{\sim \chi^2_n} = \sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum \left(\frac{\bar{X} - \mu}{\sigma}\right)^2$$

$$= \sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + n\left(\frac{\bar{X} - \mu}{\sigma}\right)^2$$

$$= \sum \frac{(X_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

By the definition of S^2 →

$$= \frac{(n-1)S^2}{\sigma^2} + \underbrace{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2}_{\sim \chi^2_1}$$

Since S^2 and $\bar{X}$ are independent, the terms in the RHS of the eq. is also independent. By the substractive prop. of χ^2 dist, we conclude that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. ■

From the previous proposition, it can be shown that

$$E(S^2) = \sigma^2 \quad \text{and} \quad V(S^2) \to 0, \text{ as } n \to \infty.$$

A t-distributed statistic

Previously, we defined t distribution as

$$T = \frac{Z}{\sqrt{V/\nu}}.$$

From this definition, it is not obvious how the t distribution can be applied to data. For this reason, we give the following result.

Gosset's theorem: If $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma)$ distribution, then the rv

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has the t distribution with $(n-1)$ degrees of freedom, t_{n-1}.

Proof: By re-expressing the fraction in a way of

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X} - \mu}{\sigma}}{\frac{S}{\sigma\sqrt{n}}} = \frac{\frac{\bar{X} - \mu}{\sigma}\sqrt{n}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}.$$

divide both numerator and denominator by σ.

multiply and divide the denominator by $(n-1)$.

Here, the numerator on the RHS is standard normal. The denominator is the square root of a χ^2_{n-1} variable, divided by its degrees of freedom. This chi-squared variable is independent of the numerator. Then, by definition the ratio has the t distribution with n-1 degrees of freedom. ■

Comparison.

Now, we have

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

↓

Z has a standard normal distribution when X_i's are iid normal rvs.

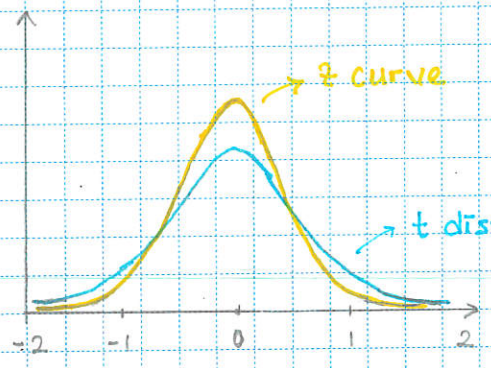
↓

Contrarily, T has a t_{n-1} distribution, when we have S instead of σ.

Here, replacing σ (constant) with S (rv) results in T having greater variability than Z. Now, recall the graph of t dist. (week 11, page 13, section 6.3)

↓

check the updated lecture notes.



→ S produces extra uncertainty.

7. Point Estimation

The objective of point estimation is to use a sample to compute a number that represents a "good guess" for the true value of the parameter. The resulting number is called a point estimate.

7.1. Concepts and criteria for point estimation

Definition. A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and determining its value from the given sample data. The selected statistic is called the point estimator of θ.

Example. Consider a random sample of $n=3$ with the observed data $x_1 = 5.0$, $x_2 = 6.4$, and $x_3 = 5.9$. Then, the sample mean $\bar{x} = 5.77$. Here point estimator is $\bar{X}$, the point estimate is $\bar{x} = 5.77$, and $\bar{x} = 5.77$ is a "best guess" for μ (true mean). Also, same estimator $\bar{X}$ can give a different estimate with different values of x_1, x_2, and x_3.

NOTATION. ⑧ $\hat{\mu} = \bar{X}$ → It is read as "the point estimator of μ is the sample mean."

⑧ $\hat{\mu}$ → We use "hat" to denote the point estimate. In the example above, $\hat{\mu} = \bar{x} = 5.77$.

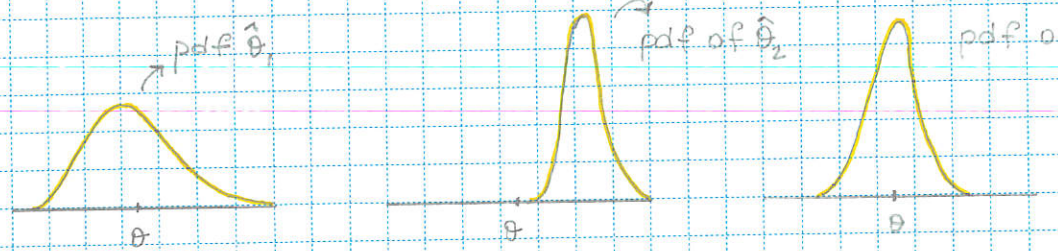
Assessing estimators: accuracy and precision

When a particular statistic is selected to estimate an unknown parameter, two criteria often used to assess the quality of that estimator: accuracy and precision. These two notions are made more rigorous by the following definitions.

Definition. A point estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. The difference $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$ (and equals 0 if $\hat{\theta}$ is unbiased).

Definition. The standard error of $\hat{\theta}$ is its standard deviation, $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the estimated standard error of $\hat{\theta}$, and it is denoted by $\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$.

→ Both bias and standard error are properties of an estimator.



- low bias
- relatively high standard error

↓

accurate, not precise

- high bias
- low standard error

↓

(The values of $\hat{\theta}_2$ across different random samples overestimate θ.)

↓

not accurate, precise

- low bias
- relatively small standard error

↓

($\hat{\theta}_3$ has a mean roughly equal to θ)

↓

accurate, precise

Mean squared error

Instead of looking at bias and variance (accuracy and precision) seperately, we can measure how close $\hat{\theta}$ is to θ using the squared error $(\hat{\theta}-\theta)^2$. An omnibus measure of quality is the mean squared error (expected squared error).

Definition. The mean squared error (MSE) of an estimator $\hat{\theta}$ is $E[(\hat{\theta}-\theta)^2]$.

The following result says that MSE gets larger if

i) the estimator is biased
ii) the estimator is highly variable

Proposition. For any estimator $\hat{\theta}$ of a parameter θ,

$$MSE = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 = \text{variance of estimator} + (\text{bias})^2$$

In particular, for any unbiased estimator of θ, its MSE and variance are equal.

...../...../.....

Unbiased estimation

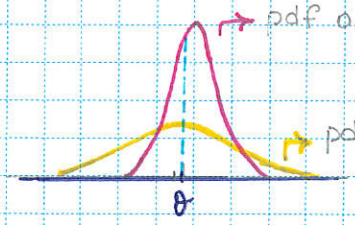
It is sometimes impossible to find an estimator with the smallest MSE for every possible value of the parameter. A common solution is to limit the set of estimators and then find the best one within that set.

Proposition. If $X_1, \dots, X_n$ is a random sample from a distribution with mean μ, then $\bar{X}$ is an unbiased estimator of μ. If the distribution is continuous and symmetric, then $\tilde{X}$ and any trimmed mean are also unbiased estimators of μ.

trimmed mean: discard the smallest and largest values of the sample and then average

Unbiasedness alone does not always identify a single estimator. When multiple unbiased estimators exist, the one with the smallest variance (or standard error) is chosen and is called the minimum variance unbiased estimators (MVUE) of θ.

Theorem. Let $X_1, \dots, X_n$ be a random sample from a normal distribution with parameters μ and σ. Then the estimator $\hat{\mu} = \bar{X}$ is the MVUE for μ.



Sometimes a slightly biased estimator is preferable

This is easier to obtain.

Consistency

Definition. Let $X_1, \dots, X_n$ be a random sample from a distribution that depends on a parameter θ. Then an estimator $\hat{\theta}$ is a consistent estimator of θ if $\hat{\theta}$ converges to θ as $n \to \infty$, either in the sense that

1. $E[(\hat{\theta} - \theta)^2] \to 0$ as $n \to \infty$ OR
2. $P(|\hat{\theta} - \theta| \geq \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$

→ Consistency is an asymptotic property.

Example. An automobile manufacturer has developed a new type of bumper, which is supposed to absorb impacts with less damage than previous bumpers. The manufacturer has used this bamper in a sequence of 25 controlled crashes against a wall, each at 10 mph, using one of its compact car models. Let X be the number of crashes that result in no visible damage to the automobile (a "success"). The parameter to be estimated is p, the proportion of all such crashes that result in no visible damage, equivalently, $p = P(\text{no visible damage in a crash})$.

1. If X is observed to be $x = 15$, find the estimator and estimate.
2. Is $\hat{P}$ unbiased?
3. Find the standard error and an upper bound on it.
4. What can we say about the precision of the estimator $\hat{P}$?
5. Find the estimated standard error when $n = 25$ and $\hat{p} = .6$.
6. Find the MSE of $\hat{P}$.
7. Can $\tilde{P} = \frac{X+2}{n+4}$ be an alternative estimator? Discuss this using MSE as the criterion.

Solution.

1. Estimator $= \hat{P} = \frac{X}{n}$ Estimate $= \hat{p} = \frac{x}{n} = \frac{15}{25} = .60$.

2. In order to see if $\hat{P}$ is unbiased, we need to check $E(\hat{P}) \stackrel{?}{=} p$, where p is the true proportion of successes in all possible crash tests. Since $X \sim Bin(n, p)$,

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} np = p.$$

Then, $\hat{P}$ is unbiased regardless of the value of p and the sample size n.

The standard error of the estimator is

$$\sigma_{\hat{p}} = \sqrt{V(\hat{p})} = \sqrt{V\left(\frac{X}{n}\right)} = \sqrt{\frac{1}{n^2} V(X)} = \sqrt{\frac{1}{n^2} np(1-p)} = \sqrt{\frac{p(1-p)}{n}}.$$

Since $\sigma_{\hat{p}}$ takes its largest value when $p = .5$, an upper bound on the standard error is $\sqrt{\frac{.5(1-.5)}{n}} = \sqrt{\frac{.25}{n}} = \sqrt{\frac{1}{4n}} = \frac{1}{2\sqrt{n}}$.

4. To answer the question regarding precision, we can look at the standard error: We can see that the standard error decreases as the sample size n increases. This means that the precision of the estimator $\hat{p}$ improves.

5. Since p is unknown, we can set $\hat{p} = \frac{x}{n}$ into $\sigma_{\hat{p}}$, which yields the estimated standard error:

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For $n = 25$ and $\hat{p} = .6$, we obtain

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{.6(1-.6)}{25}} = \sqrt{\frac{.24}{25}} = .098$$

6. We know that MSE is given by $E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$. Then,

$$E[(\hat{p} - p)^2] = V(\hat{p}) + 0^2 = \frac{p(1-p)}{n} \quad \ldots (1)$$

7. $\tilde{p} = \frac{X+2}{n+4}$ means that add two successes and two failures to the sample and then calculate the new sample proportion of successes.

Find $E(\tilde{p})$ and $V(\tilde{p})$, then apply the definition of MSE.

$$E(\tilde{p}) = E\left(\frac{X+2}{n+4}\right) = \frac{E(X)+2}{n+4} = \frac{np+2}{n+4}, \quad V(\tilde{p}) = V\left(\frac{X+2}{n+4}\right) = \frac{V(X+2)}{(n+4)^2} = \frac{V(X)}{(n+4)^2} = \frac{np(1-p)}{(n+4)^2} = \frac{p(1-p)}{n+8+16/n}$$

Thus,

$$MSE = \frac{p(1-p)}{n+8+16/n} + \left(\frac{np+2}{n+4} - p\right)^2 = \frac{p(1-p)}{n+8+16/n} + \left(\frac{2/n + 4p/n}{1+4/n}\right)^2 \quad \ldots (2)$$

If one MSE were smaller than the other for all values of p, then we could say that one estimator is always prefered to the other. But this is not true for different sample sizes. For small n, (2) is better with p is near .5. For large n, MSEs that are given by (1) and (2) are quite similar.

Some complications

$\bar{X}$ is the MVUE for a population mean when sampling from a normal distribution. However this does not mean that $\bar{X}$ should be used irrespective of the distribution being sampled.

Example. Imagine that the population distribution is a member of one of the following three families:

(1) ... $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\theta)^2}{2\sigma^2}}$, $-\infty < x < \infty$ → Normal dist.

(2) ... $f(x) = \frac{1}{\pi[1+(x-\theta)^2]}$, $-\infty < x < \infty$ → Cauchy dist.

(3) ... $f(x) = \frac{1}{2c}$, $\theta - c \leq x \leq \theta + c$ → Uniform dist.

The distributions (1), (2), and (3) are symmetric about θ. Then, θ is the median.

Consider four estimators $\bar{X}$, $\tilde{X}$, $\bar{X}_e$ (the average of the two extreme observations), and $\bar{X}_{tr}$ (a trimmed mean). The best estimator for θ depends on which distribution is being sampled:

1. If the random sample comes from a normal dist., then $\bar{X}$ is the best of the four estimators, since it has minimum variance among all unbiased estimators.

2. If the rand. sample comes from a Cauchy dist., then $\bar{X}$ and $\bar{X}_e$ are bad (they are very sensitive to outlying observations), $\tilde{X}$ is quite good.

3. If the rand. sample comes from dist. (3), then $\bar{X}_e$ is the best estimator.

4. The $\bar{X}_{tr}$ works reasonably well in all three. For this reason, a trimmed mean with small trimming percentage is said to be a robust estimator.

$$\bar{X}_e = \frac{\min(X_i) + \max(X_i)}{2}$$

Censoring. In some experiments the components are left in operation only until the time of the rth failure, where $r \leq n$. This procedure is referred to as censoring. (n is the number of components.)

Example. Suppose a type of component has a lifetime distribution that is exponential with parameter λ, so that expected lifetime is $\mu = 1/\lambda$. A sample of n such components is selected, and each is put into operation. The experiment is terminated at the time of the rth failure, where $r \leq n$. Let T_r denote the total accumulated lifetime at termination. Show that $\hat{\mu} = T_r / r$ is an unbiased estimator of μ.

Solution. Let Y_1 denote the time of the first failure, Y_2 denote the time of the second failure, and so on. Since the experiment terminates at time Y_r, the total accumulated lifetime at termination, T_r, is

$$T_r = \sum_{i=1}^{r} Y_i + (n-r) Y_r .$$

Now, recall two properties of exponential distribution:

1. The memoryless property: At any time point, remaining lifetime has the same exponential distribution as original lifetime.

2. If $X_1, \ldots, X_k$ are independent exponential rvs with parameter λ, then $\min(X_1, \ldots, X_k)$ is exponential with parameter $k\lambda$ and has expected value $1/(k\lambda)$.

T_r can also be written as

$$T_r = n Y_1 + (n-1)(Y_2 - Y_1) + (n-2)(Y_3 - Y_2) + \ldots + (n-r+1)(Y_r - Y_{r-1}).$$

n components last until Y_1 — $(n-1)$ compon. last until $(Y_2 - Y_1)$ — ...

→ NEXT PAGE

By the memoryless property,

$$E(T_r) = n E(Y_1) + (n-1) E(Y_2 - Y_1) + \cdots + (n-r+1) E(Y_r - Y_{r-1})$$

$$= n \frac{1}{n\lambda} + (n-1) \frac{1}{(n-1)\lambda} + \cdots + (n-r+1) \frac{1}{(n-r+1)\lambda}$$

$$= r \cdot \frac{1}{\lambda}.$$

Now, we have

$$E\left(\frac{T_r}{r}\right) = \frac{1}{r} E(T_r) = \frac{1}{r} \frac{r}{\lambda} = \frac{1}{\lambda} = \mu.$$

Since we showed that

$$E\left[\frac{T_r}{r}\right] = \mu \quad \left[E(\hat{\theta}) - \theta = 0\right]$$

we conclude that $\hat{\mu} = \frac{T_r}{r}$ is an unbiased estimator for μ.

Parametric bootstrap: Sometimes we have an estimate of θ but no measure of the uncertainty in that estimate. In such situations, repeated values from the pdf $f(x; \hat{\theta})$ are simulated, and the estimated standard error is obtained. This procedure is known as the parametric bootstrap.
