Distributions and Estimates

The purpose of both of these distributions is to allow for inferences about μ and σ in an unknown distribution. Both are quotients of known distributions.

Preliminaries

Sample Mean: Let $Y_1, \ldots, Y_n$ be a random, independent sample from a distribution with mean μ and variance σ^2. Then,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{Sample Mean}$$

is a distribution with mean $\overline{\mu} = \mu$ and variance $\overline{\sigma}^2 = \frac{\sigma^2}{n}$. If the underlying distribution is a normal distribution, then $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ is a *standard* normal distribution.

Sample Variance: The *sample variance* is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2. \qquad \text{Sample Variance}$$

It is important to note that the sample variance is found for samples drawn from a distribution; for population standard deviation/variance, we use n instead of $n-1$ in the denominator.

When Y_i is a normal distribution, then $\frac{(n-1)S^2}{\sigma^2}$ is a χ^2 distribution with $n-1$ df — S^2 and $\overline{Y}$ are independent.

Definition of T Distribution

Let Z be a standard normal distribution, W be χ^2 with ν df, and Z and W be independent. Then,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

has a T distribution with ν df.

Creating a T Distribution: Let Y_i be sampled from a normal distribution with mean μ and standard deviation σ.

Then, $Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ is a standard normal distribution, and $W = \frac{(n-1)S^2}{\sigma^2}$ is χ^2 with $n-1$ df.

So,

$$\begin{aligned} T &= \frac{Z}{\sqrt{W/(n-1)}} \\ &= \frac{(\overline{Y}-\mu)\sqrt{n}}{\sigma}\sqrt{\frac{(n-1)\sigma^2}{S^2}} \\ &= \frac{(\overline{Y}-\mu)\sqrt{n}}{S} \end{aligned}$$

has a T distribution with $n-1$ df.

T Distribution: Let $Y_1, \ldots, Y_6$ be samples from a normal distribution with unknown μ, σ. Estimate $P(|\overline{Y}-\mu| < (2S/\sqrt{n}))$.

Thus, we have

$$\begin{aligned} P\left(|\overline{Y}-\mu| \leq \frac{2S}{\sqrt{n}}\right) &= P\left(-2 \leq \frac{\sqrt{n}(\overline{Y}-\mu)}{S} \leq 2\right) \\ &= P(-2 \leq T \leq 2) \end{aligned}$$

Thus, for $n = 6$, we have that our random variable T has 5 df. By looking at a T distribution table, we can find that $P \approx 0.9$. We can also use R.

Definition of F Distribution

Let W_1 and W_2 be independent χ^2 distributions with ν_1 and ν_2 df respectively. Then, the F distribution with ν_1 numerator df and ν_2 denominator df is found as follows:

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

Simplifying an F Distribution: Let n_1 samples be drawn from normal distribution with mean μ_1 and variance σ_1^2, and n_2 samples be drawn from normal distribution with mean μ_2 and variance σ_2^2. Both distributions are independent.

From each of these samples, we find the sample variance, and create χ^2 distributions with their respective df.

$$\begin{aligned} W_1 &= \frac{(n_1-1)S_1^2}{\sigma_1^2} \\ W_2 &= \frac{(n_2-2)S_2^2}{\sigma_2^2} \end{aligned}$$

Therefore, we have

$$\begin{aligned} F &= \frac{W_1/(n_1-1)}{W_2/(n_2-1)} \\ &= \frac{(n_1-1)S_1^2}{\sigma_1^2(n_1-1)} \frac{\sigma_2^2(n_2-1)}{(n_2-1)S_2^2} \\ &= \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \end{aligned}$$

as an F distribution with $n_1 - 1$ numerator df and $n_2 - 1$ denominator df.

Applying the F Distribution: Let $n_1 = 6$ and $n_2 = 10$ be two samples from independent normal distributions with the same σ^2. Find b such that $P\left(\frac{S_1^2}{S_2^2} \leq b\right) = 0.95$.

$$\frac{S_1^2}{S_2^2} = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2}$$

The given F distribution has 5 numerator df and 9 denominator df. Therefore, we want to find $0.95 = P(F_{5,9} < b)$, or find the 0.95 quantile; in R, we find this with the `qt` function.

Normal Approximation of Binomial

Recall that a binomial distribution Y with n trials and p probability of success has probabilities found below:

$$P(Y \leq \ell) = \sum_{k=0}^{\ell} \binom{n}{k} p^k (1-p)^{n-k}.$$

For very large n, this sum is hard to calculate. We could approximate with the Poisson distribution, but this still requires a lot of calculations and large factorial values. Instead, we will try the following:

$$\begin{aligned} X_i &= \begin{cases} 1 & i \text{ trial success} \\ 0 & i \text{ trial failure} \end{cases} \\ E(X_i) &= p \\ E(X_i^2) &= p \\ V(X_i) &= p(1-p) \\ \overline{X} &= \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{Y}{n} \\ E(\overline{X}) &= p \\ V(\overline{X}) &= \frac{p(1-p)}{n} \end{aligned}$$

By the Central Limit Theorem, we approximate $\overline{X}$ as a normal distribution with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Alternatively, we can create, for large fixed n, $Y = n\overline{X}$ with mean np and standard deviation $\sqrt{np(1-p)}$.

For example, consider $p = 0.5$, $n = 100$, $Y =$ number of successes. To find $P\left(\frac{Y}{n} > 0.55\right)$. By the Central Limit Theorem, this is approximately a normal distribution with mean 0.5 and standard deviation 0.05.

Applying Central Limit Theorem: Let Y be a binomial distribution with $n = 25$ and $p = 0.4$. Then, $\mu = np = 10$, and standard deviation $\sigma = \sqrt{\frac{p(1-p)}{n}} = 5\sqrt{0.24}$.

To find $P(Y \leq 8)$, we can potentially approximate with $P(X \leq 8.5)$ — the reason we use 8.5 instead of 8 is due to the fact that n may not be large enough, a process known as the continuity correction.

Using standardization (or R), we find that this probability is approximately 0.269.

The actual probability $P(Y \leq 8)$ is found as below:

$$\begin{aligned} P(Y \leq 8) &= \sum_{k=0}^{8} \binom{25}{k} (0.4)^k (0.6)^{1-k} \\ &= 0.274 \end{aligned}$$

The normal approximation for the binomial is adequate when $p \pm 3\sqrt{\frac{p(1-p)}{n}} \in (0,1)$. Essentially, the binomial trial needs to have an adequate sample size such that the "spread" is small. This is equivalent to $n \geq 9\frac{\max(p,1-p)}{\min(p,1-p)}$.

Estimators

Let Y be a random variable with an *unknown* distribution.

Parameter: Feature of Y's distribution that are not computable from samples.

Examples of Parameters: μ, σ, m'_k, interval $(a, b) \ni P(y \in I) = 0.95$.

Statistic: Random variable that is computable from samples.

Examples of Statistics: sample mean, $\overline{Y}$, sample variance, S^2, $Y_{(i)}$.

Estimator: a statistic intended to approximate a parameter. A point estimator estimates a single value.

Examples of Estimators: $\overline{Y}$ as an estimator for μ, and S^2 as an estimator of σ^2.

Bias and Mean Square Error of Estimators

We want to find θ, a constant parameter of the underlying distribution — $\hat{\theta}$ is a random variable.

If $E(\hat{\theta})$ is close to θ, we can say that $\hat{\theta}$ is a good estimator — more precisely, we define the bias $B(\hat{\theta}) = E(\hat{\theta}) - \theta$, and if $B(\hat{\theta}) = 0$, then $\hat{\theta}$ is an unbiased estimator.

In addition to minimizing bias, to see whether or not an estimator is good requires minimizing the variance of the estimator — the mean squared estimator $\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$. Notice that for an *unbiased* estimator, $\text{MSE}(\hat{\theta}) = V(\hat{\theta})$.

Exercise 8.12: Let θ be the true voltage of some electronic device. The voltage test has results uniformly distributed over $[\theta, \theta + 1]$. There are n tests, $Y_1, \ldots, Y_n$. Evaluate $\overline{Y}$ as an estimator for θ.

Solution: Since the voltage is uniformly distributed over $[\theta, \theta + 1]$, we have that Y_i is uniform on $[\theta, \theta + 1]$. Therefore, $E(Y_i) = \theta + 0.5$, and $V(Y_i) = \frac{1}{12}$.

Therefore, $E(\overline{Y}) = \theta + 0.5$, and $V(\overline{Y}) = \frac{1}{12n}$, meaning $\text{MSE}(\hat{\theta}) = \frac{1}{12n} + \frac{1}{4}$.

If we want an unbiased estimator for θ, we take $\hat{\theta} = \overline{Y} - \frac{1}{2}$. Then, $E(\hat{\theta}) = E(\overline{Y}) - E(1/2) - \theta = 0$. By shifting this estimator, our new MSE is $\frac{1}{12n}$.

Example 8.1: We will compare the two estimators of σ^2: sample variance and population variance.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Solution: Recall $V(X) = E(X^2) - (E(X))^2$. Therefore, $E(X^2) = V(X) + (E(X))^2$.

$$\begin{aligned} E(Y_i^2) &= V(Y_i) + (E(Y_i))^2 \\ &= \sigma^2 + \mu^2 \\ E(\overline{Y}^2) &= V(\overline{Y}) + \left(E(\overline{Y})\right)^2 \\ &= \frac{\sigma^2}{n} + \mu^2 \end{aligned}$$

Notice that

$$\begin{aligned}
\sum(Y_i-\overline{Y})^2 &= \sum(Y_i^2-2Y_i\overline{Y}+\overline{Y}^2)\\
&= \sum Y_i^2 - 2\overline{Y}\sum Y_i + \sum \overline{Y}^2\\
&= \sum Y_i^2 - 2n\overline{Y}^2 + n\overline{Y}^2\\
&= \sum_{Y_i}^{2} -n\overline{Y}^2\\
E\left(\sum(Y_i-\overline{Y})^2\right) &= E(\sum Y_i^2) - nE(\overline{Y}^2)\\
&= n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\\
&= (n-1)\sigma^2\\
B(S'^2) &= \frac{1}{n}(n-1)\sigma^2-\sigma^2\\
&= -\frac{1}{n}\sigma^2 \neq 0\\
B(S^2) &= \frac{1}{n--1}(n-1)\sigma^2-\sigma^2\\
&= 0
\end{aligned}$$

S'^2 is known as the *biased sample variance*, while S^2 is the unbiased sample variance.

The standard error $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. If $\hat{\theta}$ is unbiased, then $\sigma_{\hat{\theta}} = \sqrt{\text{MSE}(\hat{\theta})}$

Errors and Confidence Intervals

Error of Estimation: The error of estimation is $\varepsilon = |\hat{\theta}-\theta|$. Notice that while θ is a fixed value, ε is a random variable.

We say $\hat{\theta}$ is a "good" estimator if there is a high probability that ε is small. Specifically, ε being small often means $\exists b$ such that $\varepsilon < b$ — alternatively, $|\hat{\theta}-\theta| < b$, meaning $\theta-b<\hat{\theta}<\theta+b$, so $\hat{\theta}\in(\theta-b,\theta+b)$.

We often set b to be $2\sigma_{\hat{\theta}}$, or $2\cdot\text{SE}(\hat{\theta})$.

When $\hat{\theta}$ is unbiased, $\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta$. So, the $2\sigma_{\hat{\theta}}$ interval about θ is the same as the $2\sigma_{\hat{\theta}}$ about $\hat{\theta}$.

Finally, $\hat{\theta}$ often, but not always, has an approximate normal distribution. Therefore, the probability that $\hat{\theta}$ is within $2\sigma_{\hat{\theta}}$ of $\mu_{\hat{\theta}}$ is approximately 0.95. Specifically, $P(|\hat{\theta}-\mu_{\hat{\theta}}|<2\sigma_{\hat{\theta}})$.

Recall that Chebyshev's Theorem states that $P(|\hat{\theta}--\mu_{\hat{\theta}}|<2\sigma_{\hat{\theta}})\geq 1-\frac{1}{2^2}=0.75$.

Example 8.3: Suppose there are two types of tire. $n_1=n_2=100$ of each type, with $Y_1=Y_2=$ miles tire lasts. $\overline{Y}_1=26400$ miles while $\overline{Y}_2=25100$ miles. $S_1^2=144000000$ and $S_2^2=196000000$.

Let's try to estimate how much longer tire 1 lasts than tire 2. First, we will use an unbiased estimator

for the mean (sample mean).

$$\begin{aligned}\mu_{Y_1-Y_2} &= \mu_{Y_1} - \mu_{Y_2} \\ &\approx \overline{Y}_1 - \overline{Y}_2 \\ &= 1300 \\ \sigma_{\overline{Y}_1-\overline{Y}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &\approx \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= 184.4\end{aligned}$$

Table 8.1

S^2 an unbiased estimator for σ^2

Therefore, the difference in the life expectancy between the types is about 1300 miles, and there is approximately probability 0.95 chance that the life expectancy is within 368.8 miles of 1300.

The interval $[1300 - 368.8, 1300 + 368.6]$ is called an interval estimator or confidence interval, expressed as $[\hat{\theta}_L, \hat{\theta}_H]$.

- $\hat{\theta}_L$: lower confidence limit, a left endpoint estimator.
- $\hat{\theta}_U$: upper confidence limit, a right endpoint estimator.

Example 8.4: One sample, Y, from exponential distribution with PDF

$$f(y) = \begin{cases} \frac{1}{\theta} e^{-y/\theta} & y \in [0, \infty) \\ 0 & y \in (-\infty, 0) \end{cases}$$

To estimate θ, we would prefer a PDF without θ. Let

$$U = \frac{Y}{\theta}$$

pivotal quantity

$$\begin{aligned}F_U(u) &= P(U \leq u) \\ &= P(Y/\theta \leq u) \\ &= F_Y(u\theta) \\ &= 1 - e^{-u}\end{aligned}$$

$$f(u) = \begin{cases} e^{-u} & u \in [0, \infty) \\ 0 & u \in (-\infty, 0) \end{cases}$$

We want a, b such that $P(a \leq \theta \leq b) = 0.9$. Pick c, d such that $P(c \leq U \leq d) = 0.9$.

By integrating, we find $c = -\ln(0.95) = 0.051$, and $d = 2.996$. Now,

$$\begin{aligned}0.9 &= P(0.051 \leq U \leq 2.996) \\ &= P(0.051 \leq Y/\theta \leq 2.996) \\ &= P(0.051/Y \leq 1/\theta \leq 2.996/ \leq) \\ &= P(Y/2.996 \leq \theta \leq Y/0.051).\end{aligned}$$

meaning that for $Y = 2$, there is a probability 0.9 that $\theta \in [0.668, 39]$.

Common Confidence Intervals

Last time, we used the method of pivots to find a confidence interval for θ:

(1) find a confidence interval for pivotal quantity,

(2) solve for θ.

Today, we will identify some parameters with the same pivotal quantity, meaning we can find a confidence interval directly instead of using the method of pivots.

The larger the sample size, the smaller the confidence interval, meaning the more accurate our approximation. However, we cannot just sample however many people as we want — we need to find the minimum number necessary to satisfy a certain confidence interval.

Table 8.1 discusses the various features we would like to suss out from a sampling distribution, such as μ, $\mu_1 - \mu_2$ from any distribution, and p, and $p_1 - p_2$ from a binomial distribution.

All estimators in Table 8.1 are unbiased and approximately normal. The latter feature is that which we will use in large-sample confidence intervals. We can let

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

then be a pivotal quantity. Z is an approximately standard normal distribution.

Since an unbiased estimator $\hat{\theta}$ for θ has an approximately normal distribution, and $\sigma_{\hat{\theta}}$ can be estimated from known quantities, we can find a confidence interval with coefficient $1 - \alpha$.

Considering Z, the standard normal (approximate) distribution, we find a central confidence interval by placing $\alpha/2$ in each tail, with the endpoints of the interval at $\pm z_{\alpha/2}$. Then, using R or a table, we find $z_{\alpha/2}$, and

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} \leq z \leq z_{\alpha/2}) \\ &= P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}\right) \\ &\vdots \\ &= P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) \end{aligned}$$

Therefore, $\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}$ is the lower confidence limit and $\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$ is the higher confidence limit.

Example 8.8 Comparing samples A and B, where $n_1 = 50$ for A, 12 fail, and $n_2 = 60$ for B, where 12 fail. We want to find a 0.98 confidence interval for $p_1 - p_2$. If this CI contains 0, then we see that A and B last approximately the same time.

Therefore,

$$\begin{aligned}
\hat{\theta} &= \hat{p}_1 - \hat{p}_2 \\
&= \frac{Y_1}{n_1} - \frac{Y_2}{n_2} \\
&= \frac{12}{50} - \frac{12}{60} \\
&= 0.04 \\
z_{\alpha/2} &= 2.33 \\
\sigma_{\hat{\theta}} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\
&\approx \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} \\
&\approx 0.079
\end{aligned}$$

Therefore, the 0.98 confidence interval for $p_1 - p_2$ is $[-0.145, 0.225]$, which contains zero, implying that we can assume A and B fail at similar rates with probability 0.98.

Selecting Sample Size

We will select n by guesstimating things. Recall that for the estimators in Table 8.1, the confidence interval for $1-\alpha$ is $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$.

Example 1: Use $\overline{Y}$ to estimate μ of yield Y tons. We want a 0.95 confidence interval with radius 5 tons. Therefore, we want to find n for this radius.

By the central limit theorem, we know that $\overline{Y}$ is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Therefore, $\overline{Y}$ is within $2\sigma_{\overline{Y}}$ with probability 0.95, meaning $z_{\alpha/2} = 2$.

We want $z_{\alpha/2}\sigma_{\hat{\theta}} = 5$. Therefore, $n = \frac{4\sigma^2}{25}$. However, we also do not know σ. If we had a sample, we could use $\sigma^2 \approx S^2$. However, we want to find n, we do not have a sample.

Suppose we have an idea of the range of Y, approximately 84. Then, $\sigma \approx 21$, as the range is approximately 4σ, so $n \approx 71$.

The results suggest that if we took 71 samples, there is a 95% confidence that the samples will be within 5 tons of the true mean. Note that $\overline{Y}$ will likely have an error much lower than 5.

(1) If this 0.95 confidence interval is a true 0.95 confidence interval, then there is a 0.68 probability that our sample mean will be within one standard deviation, or an error of 2.5 tons.

(2) There is a probability that σ is *smaller* than 1/4 of the range.

Example 2: Suppose there is a new drug with effects A or B. We will use the sample probability to estimate the real probability of A, $\hat{p} = \frac{Y}{n}$. To get the drug approved by the FDA, we require a 0.90 confidence interval of the effect being within 0.04. We want to find a valid clinical trial size if $p \approx 0.6$, or if we have no knowledge of p.

Since we want a 0.9 confidence interval, we find $z_{\alpha/2} = z_{0.05} = 1.645$. Then, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, so $1.645\sigma_{\hat{p}} = 0.04$, so $n = 406$.

For p unknown, we have $p \approx 0.05$ to maximize $p(1-p)$, so there $n = 423$.

Example 3: We have two methods of training, creatively named method 1 and method 2, with n_1 and n_2 trainees in each method respectively. We want a 0.95 confidence interval that the difference between $\overline{Y}_1$ and $\overline{Y}_2$, the sample mean of the average times, is within 1 minute of the true mean. We expect the range of Y_1 and Y_2 to be about 8 minutes.

We have $1 - \alpha = 0.95$, meaning $z_{\alpha/2} = z_{0.025} = 1.96$. We find

$$\begin{aligned} 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}} &= 1 \\ \sigma_1^2 = \sigma_2^2 &= \sigma^2 \\ \sigma &\approx \frac{8}{4} \\ n_1 = n_2 &= n \end{aligned}$$

Solving for n, we find $n \approx 31$.

Small Sample Confidence Intervals

Recall that for an unbiased, approximately normal estimator, $\hat{\theta}$, for the parameter θ, we have a pivotal quantity $Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ is approximately standard normal with confidence interval endpoints $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$ for confidence coefficient $1-\alpha$.

However, there are some possible problems.

(1) If the sample size is small, then $\hat{\theta}$ may not be approximately normal. For example, $\hat{\theta} = \overline{Y}$ when Y is not normal.

(2) $\sigma_{\hat{\theta}}$ can have the unknown σ in it, which has to be estimated with S.

To estimate μ, we use

$$\begin{aligned} T &= \frac{Z}{\sqrt{W/\nu}} \\ &= \frac{\overline{Y} - \mu}{S/\sqrt{n}}. \end{aligned}$$

This leads to us using

$$\overline{Y} + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

for our confidence interval.

Example 2: Let $n = 8$ for a sample of muzzle velocities. Then, we get $\overline{Y} = 2959$, with $S = 39.1$. We assume the muzzle velocities are distributed approximately normally. We want to find a 0.95 confidence interval for μ.

Then, we have $t_{\alpha/2} = t_{0.025}$, with 7 df. Therefore, we find our 0.95 confidence interval as $(2927, 2996)$.

If we use z-scores instead of t-scores, we would get a confidence interval of $(2932, 2986)$, which is a narrower confidence interval than the t scores.

As a rule of thumb for estimating μ or $\mu_1 - \mu_2$ confidence intervals, use z if we either know σ or n is large enough such that the central limit theorem is effective (at least 30). We use t if σ is unknown *and* $n < 30$.

Example 2: We will build a t distribution pivot for $\mu_1 - \mu_2$.

$$z = \frac{}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

We will assume that $\sigma_1 = \sigma_2$. Then,

$$= \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

$$w = \frac{(n-1)S^2}{\sigma^2}.$$

We ask whether S_1^2 or S_2^2 should be used for sample variance. The answer is both — by taking a weighted average.

The pooled sample variance estimator is

$$S_p^2 = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \overline{Y}_2)}{n_1 + -1 + n_2 - 1},$$

so we can have

$$w = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$$

as our χ square distribution with $n_1 + n_2 - 2$ df. We make our new T distribution,

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which serves as the pivotal quantity for $\mu_1 - \mu_2$, with confidence intervals $(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Confidence Intervals for σ^2

The unbiased estimator for σ^2 is S^2. Recall for any sample $Y_1, \ldots, Y_n$,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

When the sample is from a normal distribution, then $\frac{(n-1)S^2}{\sigma^2}$ is our pivotal quantity that has a χ^2 distribution with $n - 1$ df.

There are two important differences between this χ^2 pivot (and associated confidence interval) with how we used t and z. For a $1 - \alpha$ confidence interval, we want a and b such that

$$1 - \alpha = P(a \leq \chi^2 \leq b).$$

When we choose z and t, their respective distributions are symmetric about 0 — using $b, -b$ results in the narrowest possible confidence interval. However, χ^2 is not symmetric about 0 — it's not even symmetric.

Therefore, we need to choose a and b to minimize the width of the confidence interval.

Instead, we will choose a and b to mimic symmetry with respect to the tails.

$$1-\alpha = P\left(\chi^2_{1-\alpha/2} \leq \chi^2 \leq \chi^2_{\alpha/2}\right)$$

Solving for σ^2,

$$= P\left(\frac{\chi^2_{1-\alpha/2}}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{\alpha/2}}{(n-1)S^2}\right)$$
$$= P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right).$$

Note: The subscripts refer to right tails, so use $1-\alpha/2$ when using the `qchisq` function.

Because of the central limit theorem, the z and t-inspired confidence intervals are still good even if the underlying distribution is not approximately normal (for n sufficiently large). Our χ^2 confidence interval does not work well if Y is not approximately normal, even if n is large.

Example: Let Y be the measurement of sound volume. Assume Y is approximately normal, with samples $Y_1 = 4.1$, $Y_2 = 5.2$, and $Y_3 = 10.2$. We want to find σ^2 with 0.90 confidence interval.

From the data, we can find S^2 as 10.57, with the $\alpha/2$ endpoint at 0.05 and $1-\alpha/2$ endpoint at 0.95 and $(n-1) = 2$ df. Therefore, our confidence interval is $(3.53, 205.24)$.

This is a very wide confidence interval.

Properties of Point Estimators

There are multiple desirable properties that we will focus on throughout this chapter.

Relative Efficiency

For an unbiased estimator $\hat{\theta}$, we want $V(\hat{\theta})$ to be minimized (recall that for an unbiased estimator, the MSE of $\hat{\theta}$ is equal to the variance).

To determine relative efficiency of two unbiased estimators for the same parameter, $\hat{\theta}_1$ and $\hat{\theta}_2$. Then, the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}.$$

Book Example 1: Suppose Y is normal. The following are unbiased estimators for the mean:

(1) Sample Mean: $\overline{Y}$

(2) Sample Median: $\hat{m}$.

For large n, the variance of the sample median is $(1.2533)^2(\sigma^2/n)$. We know that $V(\overline{Y}) = \sigma^2/n$. Therefore,

$$\text{eff}(\hat{m}, \overline{Y}) = 0.6366 < 1,$$

so $\overline{Y}$ is a better estimator of population mean.

Book Example 2: Let Y be uniform on $[0, \theta]$. We have two unbiased estimators, $\hat{\theta}_1 = 2\overline{Y}$ and $\hat{\theta}_2 = \frac{n}{n+1}Y_{(n)}$.

By calculating the relative efficiency, we find $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = 3/(n+2)$. Therefore, $\hat{\theta}_2$ is a much more efficient estimator for large n.

Example 3: If Y is an unknown distribution, we have $\overline{Y}$ as a sample mean, and $\overline{Y}_{n-1}$, the average of the first $n-1$ data points, which are both unbiased estimators of μ. We can see that $\overline{Y}$ is a better estimator than $\overline{Y}_{n-1}$.

Consistency

Suppose we flip a coin infinitely many times — we have a fair coin, so the probability of success is $p = 1/2$. Let Y_n be the number of heads after n flips.

This yields us an infinite sequence of random variables — let $\frac{Y_n}{n}$ be the nth approximation of p, denoted $\hat{p}_n$. If we graph $\hat{p}_n$ as a sequence, we find that as n gets large, $\hat{p}_n$ approaches p. In limit language,

$$\lim_{n\to\infty} \hat{p}_n = p.$$

However, since $\hat{p}_n$ is a function and we don't have a proper definition of convergence we will use a different formulation:

$$\lim_{n\to\infty} P(|\hat{p}_n - p| < \varepsilon) = 1 \qquad \forall \varepsilon > 0.$$

This is the definition of a weakly consistent estimator. We say that $\hat{\theta}_n$ converges in probability to θ if this is the case.

Chebyshev's Theorem: For any random variable Y whose μ and σ exist, and for any $k > 0$,

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$
$$P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Theorem 1: If $\hat{\theta}_n$ is an *unbiased* estimator of θ for each n, and $\lim_{n\to\infty} V(\hat{\theta}_n) = 0$, then $\hat{\theta}_n$ is consistent. Notice that there must be two things that exist for this to exist: finite variance and a finite expected value.

Proof: Let $\varepsilon > 0$. Since $\hat{\theta}_n$ is an unbiased estimator, $E(\hat{\theta}_n)$, and since we are assuming the variance exists for each $\hat{\theta}_n$, then $\sigma_{\hat{\theta}_n}$.

Suppose $\sigma_{\hat{\theta}_n} \neq 0$. For each n, set $k_n = \frac{\varepsilon}{\sigma_{\hat{\theta}_n}}$. Applying Chebyshev's theorem, we have

$$P(|\hat{\theta}_n - E(\hat{\theta}_n)| \geq k_n\sigma_{\hat{\theta}_n}) \leq \frac{1}{k_n^2}.$$
$$P\left(|\hat{\theta}_n - \theta| \geq \frac{\varepsilon}{\sigma_{\hat{\theta}_n}}\sigma_{\hat{\theta}_n}\right) = \frac{\sigma_{\hat{\theta}_n}^2}{\varepsilon^2}$$
$$0 \leq P(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq \frac{V(\hat{\theta})}{\varepsilon}$$
$$\lim_{n\to\infty} 0 \leq \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{V(\hat{\theta})}{\varepsilon}$$
$$0 \leq P(|\hat{\theta}_n - \theta| \geq \varepsilon) \leq 0$$

Therefore, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon) = 0$, meaning $\hat{\theta}_n$ is consistent.

Sample Mean: $\overline{Y}_n$ is an unbiased estimator for μ, and

$$\begin{aligned}\lim_{n\to\infty} V(\overline{Y}_n) &= \lim_{n\to\infty} \frac{\sigma^2}{n} \\ &= 0.\end{aligned}$$

Therefore, $\overline{Y}_n$ is consistent.

Sample Probability: $\hat{p}_n$ is an unbiased estimator for p.

$$\begin{aligned}\lim_{n\to\infty} \hat{p}_n &= \lim_{n\to\infty} \frac{p(1-p)}{n} \\ &= 0.\end{aligned}$$

Therefore, $\hat{p}_n$ is also consistent.

Theorem 2: If $\hat{\theta}_n$ and $\hat{\varphi}_n$ converge in probability to θ and φ, then

(a) $\hat{\theta}_n + \hat{\psi}_n$ converges in probability to $\theta + \psi$

(b) $\hat{\theta}_n\hat{\psi}_n$ converges in probability to $\theta\psi$

(c) $\frac{\hat{\theta}_n}{\hat{\psi}_n}$ converges in probability to $\frac{\theta}{\psi}$ so long as ψ and $\hat{\psi}_n$ do not equal zero at any point.

(d) For any f continuous at θ, $f(\hat{\theta}_n)$ converges in probability to $f(\theta)$.

Proof (a): Let $\varepsilon, \varepsilon_1 > 0$. Since $\hat{\theta}_n \to \theta$ in probability, so $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \geq \varepsilon/2) = 0$ Therefore, $\exists N_\theta \in \mathbb{Z}^+$ such that for all $n > N_\theta$, $|P(|\hat{\theta}_n - \theta| \geq \varepsilon)| < \varepsilon_1/2$. Similarly, $\exists N_\psi$ such that for all $n > N_\psi$, $|P(|\hat{\psi}_n - \psi| \geq \varepsilon)| < \varepsilon_1/2$. Let $N = \max\{N_\theta, N_\psi\}$. Then,

$$\begin{aligned}P(|\left(\hat{\theta}_n + \hat{\psi}_n\right) - (\theta + \psi)| \geq \varepsilon) &= P(|(\hat{\theta}_n - \theta) + (\hat{\psi} - \psi)| \geq \varepsilon) \\ &\leq P(|\hat{\theta}_n - \theta| + |\hat{\psi}_n - \psi| \geq \varepsilon) \\ &\leq P(|\hat{\theta}_n - \theta| \geq \varepsilon/2 \text{ or } |\hat{\psi}_n - \psi| \geq \varepsilon/2) \\ &\leq P(|\hat{\theta} - \theta| \geq \varepsilon/2) + P(|\hat{\psi}_n - \psi| \geq \varepsilon/2) \\ &< \varepsilon_1\end{aligned}$$

Sample Means: From the above theorem, $\overline{Y}_1 - \overline{Y}_2$ is a consistent estimator for $\mu_1 - \mu_2$.

Sample Probabilities: From the above theorem, $\hat{p}_1 - \hat{p}_2$ is a consistent estimator for $p_1 - p_2$.

Theorem 3: If $Y_1, Y_2, \ldots, Y_n$ represent a random sample such that μ_4', μ_2', and μ exist, then

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$$

is a consistent estimator for σ^2.

Proof:

$$\begin{aligned}S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}_n^2\right) \\ &= \underbrace{\frac{n}{n-1}}_{(3)}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}_{(1)} - \underbrace{\overline{Y}_n^2}_{(2)}\right).\end{aligned}$$

Concerning (1), we see that it is an average of samples from Y^2. Therefore,

$$\begin{aligned} V(Y^2) &= E(Y^4) - (E(Y^2))^2 \\ &= \mu_4' - (\mu_2')^2 < \infty. \end{aligned}$$

We know from above that $V(\overline{Y}^2)$ is consistent, meaning (1) is consistent for μ_2'.

Concerning (2), we can see from Theorem 2(d) that $\overline{Y}_n^2$ converges in probability to μ^2. By Theorem 2(a), we can see that $(1) - (2)$ converges in probability to $\mu_2' - \mu^2 = \sigma^2$.

Therefore, since $\frac{n}{n-1} \to 1$, S_n^2 converges in probability to σ^2

Theorem 9.3: If U_n has a distribution function that converges to a standard normal distribution, and W_n converges in probability to 1, then the distribution function of $\frac{U_n}{W_n}$ converges to the standard normal distribution.

Corollary: Since S_n^2 converges in probability to σ^2, then $\overline{Y} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ is a good confidence interval by the central limit theorem, even if σ^2 is approximated with S^2.

Sufficiency

Intuition: Estimator "uses" all the predictive power of the sample. Sufficient estimators can give a method of finding low-variance estimators.

Definition: Let $\hat{\theta}$ be a statistic; the statistic is sufficient for θ if the conditional joint distribution of $Y_1, \ldots, Y_n$ given $\hat{\theta}$ is not dependent on θ.

The discrete case is $P(y_1, \ldots, y_n|\hat{\theta})$, and the continuous case is $f(y_1, \ldots, y_n|\hat{\theta})$.

If $\hat{\theta}$ is sufficient, then any injective function of $\hat{\theta}$ is also a sufficient statistic for θ.

Example (∗): Consider a binomial distribution with parameter p. Consider

$$Y_i = \begin{cases} 1 & i\text{th trial success} \\ 0 & \text{otherwise} \end{cases}.$$

Consider $Y = \sum Y_i$ for p.

$$\begin{aligned} P(y_1, \ldots, y_n|y) &= \frac{P(Y_2 = y_1 \cap Y_2 = y_2 \cap \cdots \cap Y_n = y_n \cap Y = y)}{P(Y = y)} \\ &= \frac{p^{y_1}(1-p)^{y_1} \cdots p^{y_n}(1-p)^{y_n}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{p^{\sum y_i}(1-p)^{n-\sum y_i}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{1}{\binom{n}{y}} \end{aligned}$$

Since the conditional probability for y does not depend on p, we show that $\sum Y_i$ is sufficient for p.

Likelihood: Let Y have a distribution with parameter θ. The likelihood of the random sample values $y_1, \ldots, y_n$ given the value of θ is θ is $L(y_1, y_2, \ldots, y_n; \theta)$

$$\begin{aligned} L(y_1, \ldots, y_n; \theta) &= P(y_1, \ldots, y_n) && \text{if discrete} \\ &= f(y_1, \ldots, y_n) && \text{if continuous} \\ &= L(\theta). \end{aligned}$$

In example $(*)$, $L(y_1, \ldots, y_n; \theta) = P(y_1, \ldots, y_n) = \prod p(y_i)$.

Factorization Theorem: Let U be a statistic based on $Y_1, \ldots, Y_n$. U is sufficient for θ if and only if

$$L(y_1, \ldots, y_n; \theta) = g(u, \theta) \times h(y_1, \ldots, y_n)$$

for non-negative functions g (exclusively a function of u and θ) and h (exclusively a function of the sample). Furthermore, whatever appears in g is probably a useful statistic.

In example $(*)$, our $g(\sum y_i, p) = p^{\sum y_i}(1-p)^{n-\sum y_i}$, meaning $h(y_1, \ldots, y_n) = 1$.

Example $(**)$: Let $Y_1, \ldots, Y_n$ be a random sample where Y_i possesses the probability density function

$$f(y) = \frac{2y}{\theta^2} \qquad \text{on } [0, \theta].$$

This is tricky, seeing as the domain of f is dependent on θ. We will modify the formula so that it works on $[0, \infty)$ regardless of θ. Let $I(y) = 1$ if $y \in [0, \theta]$ and $I(y) = 0$ otherwise.

Then, $f(y) = \frac{2y}{\theta^2} I(y)$ works on $[0, \infty)$. Then,

$$\begin{aligned} L(\theta) &= L(y_1, \ldots, y_n; \theta) \\ &= \prod f(y_i) \\ &= \prod \frac{2y_i}{\theta^2} I(y_i) \\ &= 2^n \frac{1}{\theta^{2n}} \prod y_i \prod I(y_i) \end{aligned}$$

Efficient Estimators

Recall relative efficiency, $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$.

The support of a function $f : A \to B$ is $\{a \in A \mid f(a) \neq 0\}$ or its closure. For example, if the support of the joint probability density function, $f(y_1, y_2)$ is not a rectangle, then Y_1 and Y_2 are dependent.

Cramér-Rao Theorem: Let Y be a distribution with parameter θ. Let $\hat{\theta}$ be an unbiased estimator for θ. If

- $f(y; \theta)$ has continuous first and second partial derivatives;
- the support of f is independent of θ.

Then,

$$\begin{aligned} V(\hat{\theta}) &\geq \left(\underbrace{nE\left[\left(\frac{\partial \ln f}{\partial \theta} \right)^2 \right]}_{\text{Fisher Information}} \right)^{-1} \\ &= \left(-nE\left[\frac{\partial^2 \ln f}{\partial \theta^2} \right] \right)^{-1} \end{aligned}$$

where n denotes the number of samples.

Example (Uniform): Cramér-Rao does not work on a uniform distribution over $[0, \theta]$.

Example (Binomial): We have the binomial distribution with parameters p and n. We know that an unbiased estimator for p is $\frac{Y}{n}$.

$$\begin{aligned}
V(Y/n) &= \frac{1}{n^2}V(Y) \\
&= \frac{p(1-p)}{n}. \\
P(y,p) &= \binom{n}{y}p^y(1-p)^{n-y} \\
\ln(P(y,p)) &= \ln\binom{n}{y} + y\ln p + (n-y)\ln(1-p) \\
\frac{\partial}{\partial p}\ln(P(y,p)) &= \frac{y}{p} - \frac{n-y}{1-p} \\
\frac{\partial^2}{\partial p^2}\ln(P(y,p)) &= -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} \\
E\left[\frac{\partial^2}{\partial p^2}\ln(P(y,p))\right] &= -\frac{1}{p^2}E(Y) - \frac{1}{(1-p)^2}(n - E(Y)) \\
&= -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} \\
&= -\frac{n}{p(1-p)}
\end{aligned}$$

Therefore, the Cramér-Rao bound is

$$\left(-n_s \cdot -n\frac{1}{1-p}\right)^{-1} = \frac{p(1-p)}{n_s n}.$$

Definition: An unbiased estimator $\hat{\theta}$

(1) is efficient if its variance equals the Cramér-Rao bound.

(2) is minimum variance (or best) if no other estimator has smaller variance.

(3) has efficiency $\text{eff}(\hat{\theta}, \text{C-R}) \leq 1$.

Some parameters do not have efficient estimators. Some parameters don't even have a best estimator.

Theorem (Rao-Blackwell): Let $\hat{\theta}$ be an unbiased estimator for θ. Let U be a sufficient statistic for θ. Define $\hat{\theta}^*$ as

$$\hat{\theta}^* = E(\hat{\theta}|U).$$

Then, $\hat{\theta}^*$ is an unbiased estimator of θ, and $V(\hat{\theta}^*) \leq V(\hat{\theta})$. Minimum variance unbiased estimators are always sufficient estimators.

The sufficient statistic that produces minimum variance unbiased estimators is often the one in the likelihood factorization.

To find a MVUE:

- Find the statistic in the factorization theorem.
- Find an unbiased function of the estimator.

Weibull Distribution: $f(y) = \frac{2y}{\theta} e^{-y^2/\theta}$ on $[0, \infty)$.

$$\begin{aligned} L(y_1, \ldots, y_n; \theta) &= \prod f(y_i) \\ &= \frac{2^n}{\theta^n} \prod y_i \cdot e^{-1/\theta \sum y_i^2}. \end{aligned}$$

We have $\sum y_i^2$ as our sufficient statistic, and $E(U) = n\theta$. Therefore, $1/n \sum y_i^2$ is our minimum variance unbiased estimator.