5 Estimation

Simple random sample

A *simple random sample (SRS) of size n* is a sample chosen so that each subset of n individuals is _____. To draw a simple random sample of size n from a population of size N,

- number individuals in population with 1 through N
- generate n distinct random integers in _____, and use the corresponding individuals

Each sample in this course is (assumed to be) a simple random sample. (Was the sample in the article you just read an SRS? If not, the conclusion may be _____.)

A note on independence in an SRS

Many theorems require the independence of items in a random sample. Two random variables are *independent* if the realization of either one does not change the distribution of the other. e.g.

- Independent RVs:
- Dependent RVs:

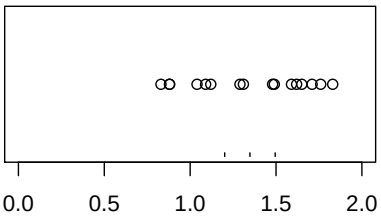
An SRS is drawn _____, that is, an item is not replaced in the population after it is selected (so it cannot be selected more than once).

Note: Items in an SRS are not independent, but they're approximately independent if the sample size is _____ relative to the population size (which it should always be in this course).

Estimating a population mean, μ

e.g. A car manufacturer uses an automatic device to paint engine blocks. Since engine blocks get hot, the paint must be heat-resistant and thin. A warehouse contains thousands of painted blocks. The manufacturer wants to know the average amount of paint applied, so 16 blocks are selected at random, and the paint thickness is measured in mil ($\frac{1}{1000}$ inch):

1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83, 1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49



Before sampling, we regard $X_1, \ldots, X_{16}$ as independent and identically distributed with _____ mean μ and _____ variance σ^2. How should we estimate μ? Estimator $\hat{\mu} =$ _____.

Note that an _____ is the formula that describes how the sample will be used to compute a guess about μ; it's a random variable. The number computed from the actual sample data is an _____, a realization of that RV. Our estimate is _____ (draw).

Theorem: If $X_1, \ldots, X_n$ are independent and identically distributed with mean μ and variance σ^2, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has mean $E(\bar{X}) =$ _____, variance $\mathrm{VAR}(\bar{X}) =$ _____, and standard deviation $\mathrm{SD}(\bar{X}) =$ _____.

Proof:

A point estimate alone isn't very useful. Reporting it with its estimated standard deviation is useful, but it's more common to report a _____ around a point estimate: coming soon.

Did our sample come from a normal distribution?

In many common situations, it is reasonable to assume that our sample is from a _____ population. This leads to a strong statement about the distribution of the sample mean:

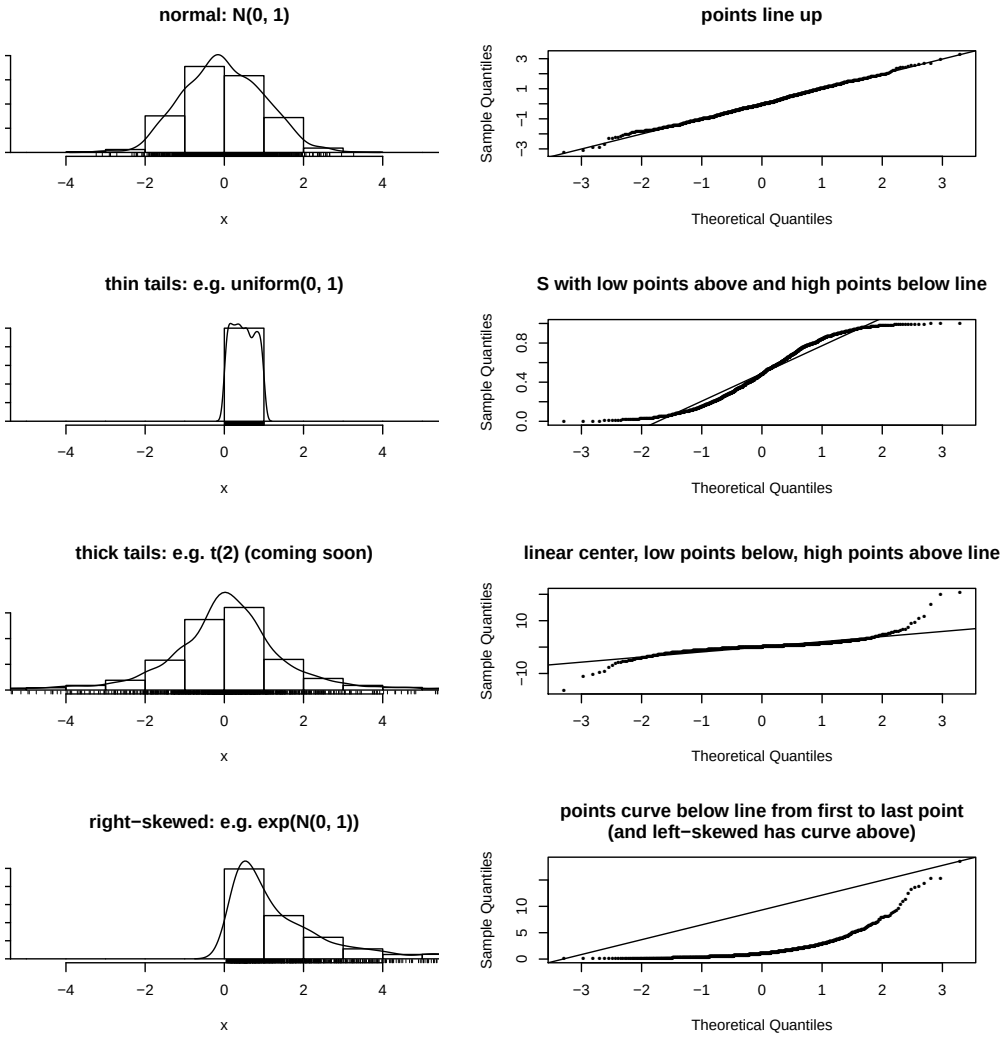
Theorem: If $X_1, \ldots, X_n$ is a simple random sample from a normal population with mean μ and variance σ^2 (so $X_i \sim N(\mu, \sigma^2)$), then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

(The proof is omitted.)

Soon we will use this theorem to derive a *confidence interval*. First, let's look at one way to assess whether a particular sample came from a normal population.

Normal probability plots

Many textbooks and statisticians use a *normal probability plot* (or *normal quantile-quantile plot* or *normal QQ plot*) to decide whether a data set is plausibly a simple random sample of size n from a _____. This plot depicts the $\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n}{n}$ quantiles from $N(0, 1^2)$ on the x-axis against the sorted data set ($\approx$ the corresponding quantiles of the population from which the sample was drawn) on the y-axis. The idea is that, if the points more-or-less _____, the data are _____ from a normal distribution. If the points do not line up, the data are _____ a normal distribution. Here are some details:



For large samples, this seems _____. For small samples, I'm _____, as I can't tell the difference between _____ in sampling and non-normality in the population.

e.g. In R, try `n = 1000; x = rnorm(n); qqnorm(x)`. Then try `n = 10` or `n = 30` many times. Also try replacing `rnorm` with `runif` (thin tails, uniform$(0, 1)$) and `rlnorm` (right-skewed, $\exp(N(0, 1))$).

The Central Limit Theorem

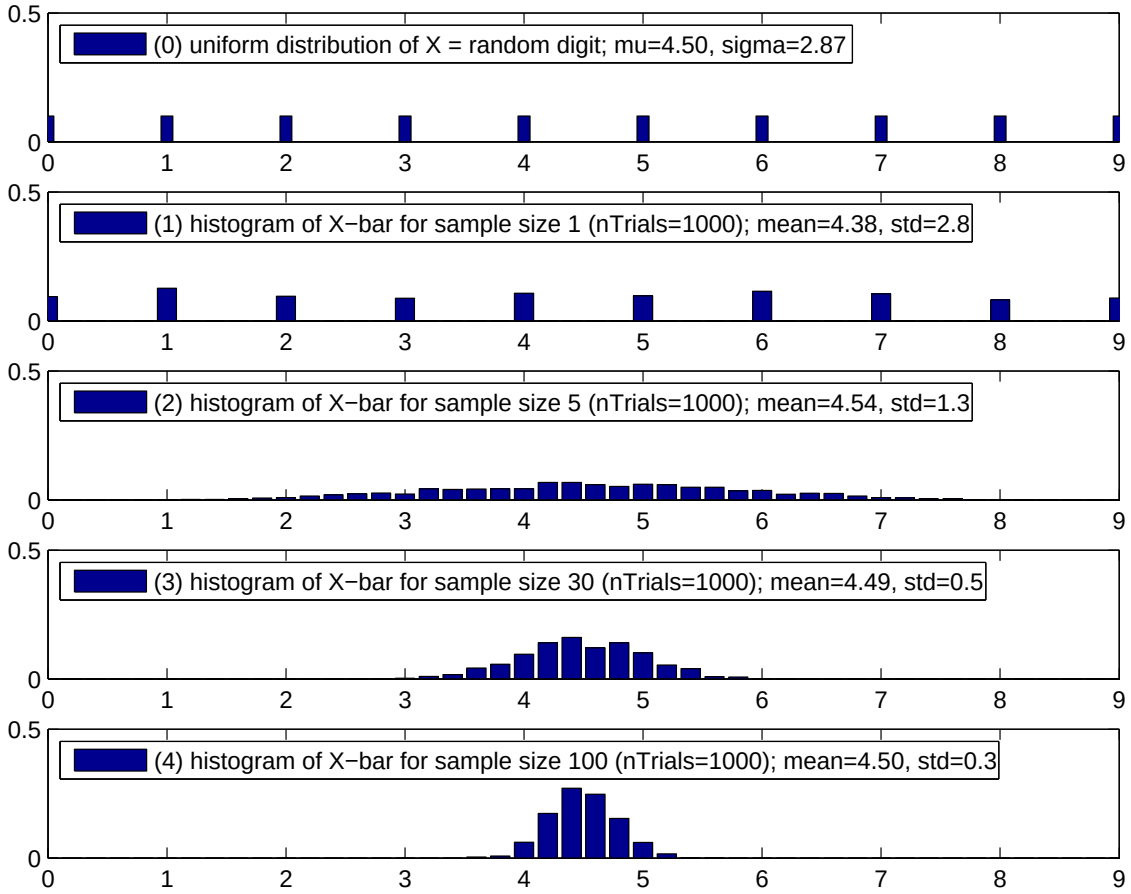
The *Central Limit Theorem* (CLT) says that the mean, $\bar{X}$, of a large enough sample from (almost) _____ distribution with finite μ and σ, is $\approx$ _____:

If $X_1, \cdots, X_n$ is a simple random sample from almost any population with finite mean μ and standard deviation σ, and n is _____, then $\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$ $(\approx)$.

(The proof is omitted. Don't miss the _____.)

($n > 30$ often counts as "large enough".)

e.g. Here is a simulation of the generation of many random samples from the discrete distribution with mass function $p(x) = \frac{1}{10}$ for $x \in \{0, 1, \cdots, 9\}$ (and 0 otherwise):



e.g. An insurance company knows that in the population of millions of homeowners, the mean annual loss from fire is $\mu = \$250$ and the standard deviation is $\sigma = \$1000$. (The loss distribution is strongly right-skewed, since most policies have no loss but a few have large losses.) If the company sells 10,000 policies, can it safely base its rates on the assumption that the average loss will be no greater than \$275?

Confidence Intervals for an Unknown Population Mean μ

We have two situations in which $\bar{X} \sim N(\mu, \sigma^2/n)$: (1) the population is $N(\mu, \sigma^2)$, for small or large n; then "$\sim$", above, is exact. (2) The sample size n is large enough that CLT applies: then "$\sim$" above is approximate. Suppose, then, that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Here we construct an interval around $\bar{X}$ which contains μ for a proportion $1-\alpha$ of random samples, where $\alpha \in (0,1)$. $100\%(1-\alpha)$ is the *confidence level* of the interval.

Let $z_{\alpha/2} = z$-score cutting off right tail area _____ from $N(0,1)$ (draw).

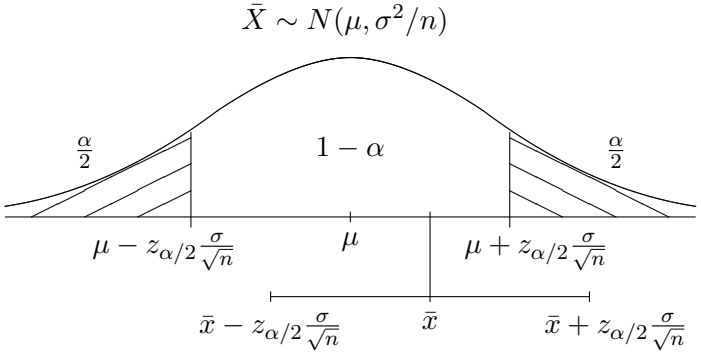
e.g. For the conventional confidence level 95%, $\alpha =$ _____ and $z_{\alpha/2} =$ _____.

Then $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ (draw). Substitute $Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$ (where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$) to get

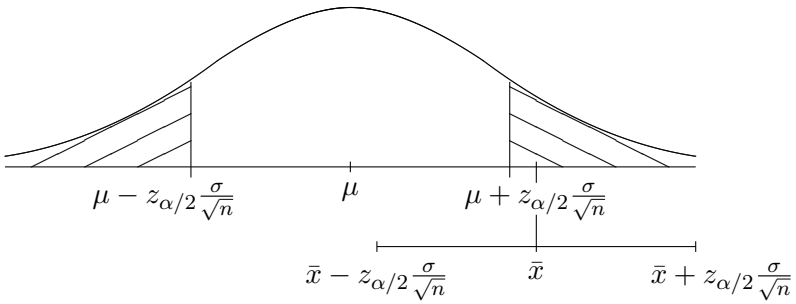
$P(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < z_{\alpha/2}) = 1-\alpha$, which we solve in two ways:

- for $\bar{X}$ in the middle: $P(\mu - z_{\alpha/2}\sigma_{\bar{X}} < \bar{X} < \mu + z_{\alpha/2}\sigma_{\bar{X}}) = 1-\alpha$ (see picture)
- for μ: $P(\bar{X} - z_{\alpha/2}\sigma_{\bar{X}} < \mu < \bar{X} + z_{\alpha/2}\sigma_{\bar{X}}) = 1-\alpha$ (see picture)

For this $\bar{x}$, μ is _____ the confidence interval. This happens with probability _____.



For this $\bar{x}$, μ is _____ the confidence interval. This happens with probability _____.



Summary:

If $X_1, \ldots, X_n$ is a simple random sample from $N(\mu, \sigma^2)$ or n is large (say $n > 30$), and σ is known, then $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ contains μ for a proportion $1 - \alpha$ of random samples. It's the $100\%(1 - \alpha)$ *confidence interval* for μ.

This form is useful when we know σ, which is _____.

e.g. Suppose we know $\sigma_{\text{paint thickness}} = 0.30$ mil. Find a 95% CI for μ.

$n =$ _____ ; Is n large enough or is sample from normal population? (Try `qqnorm(paint)`.)

$1 - \alpha =$ _____ $\implies \alpha =$ _____ $\implies z_{\alpha/2} =$ _____

$\bar{x} =$ _____, error margin = _____

$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} =$ _____

With what probability does our interval contain μ? _____

How Confidence Intervals Behave

- $\bar{X} \pm$ _____$\frac{\sigma}{\sqrt{n}}$ is a 68% confidence interval for μ
- $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is a 95% confidence interval for μ (and $1.96 \approx$ _____)
- $\bar{X} \pm$ _____$\frac{\sigma}{\sqrt{n}}$ is a 99% confidence interval for μ
- $\bar{X} \pm$ _____$\frac{\sigma}{\sqrt{n}}$ is a 99.7% confidence interval for μ

We want high confidence and a small margin of error, but the margin is $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, which gets smaller when $z_{\alpha/2}$ gets smaller, which corresponds to $(1 - \alpha)$ getting smaller too. Extreme cases are that we can have confidence approaching 100% as the margin approaches _____, or we can have confidence approaching _____ as the margin approaches 0.

Choosing the Sample Size

Good news is that the margin also gets smaller as _____. For a desired margin of error m, we can find the required sample size:

$m = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies$

(Use $\sigma \approx s$ in the usual case where we don't know σ.)

e.g. What sample size is required to reduce the error margin of the paint thickness 95% confidence interval, above, to 0.1 mil?