Sampling Distributions

Sample Summaries

Population

- Size N (usually ∞)
- Mean = μ

$$\mu = \sum p_j X_j \quad or \quad \int \dots$$

- Variance = σ^2

$$\sigma^2 = \sum p_j \left(X_j - \mu\right)^2 \quad or \quad \int \dots$$

Sample

- Size n
- Mean = $\overline{X}$

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

- Sample variance = s^2

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2$$

Sums of Normal Random Variables

We already know that linear functions of a normal rv are normal. What about combinations (eg. sums) of normals?

==> If $X_j \sim N\left(\mu_j, \sigma_j^2\right)$ (indep) then

$$Y = \sum_{j=1}^{n} X_j$$

$$Y \sim N\left(\sum_{j=1}^{n} \mu_j, \sum_{j=1}^{n} \sigma_j^2\right)$$

Combine this with what we have learned about linear functions of means and variances to get ...

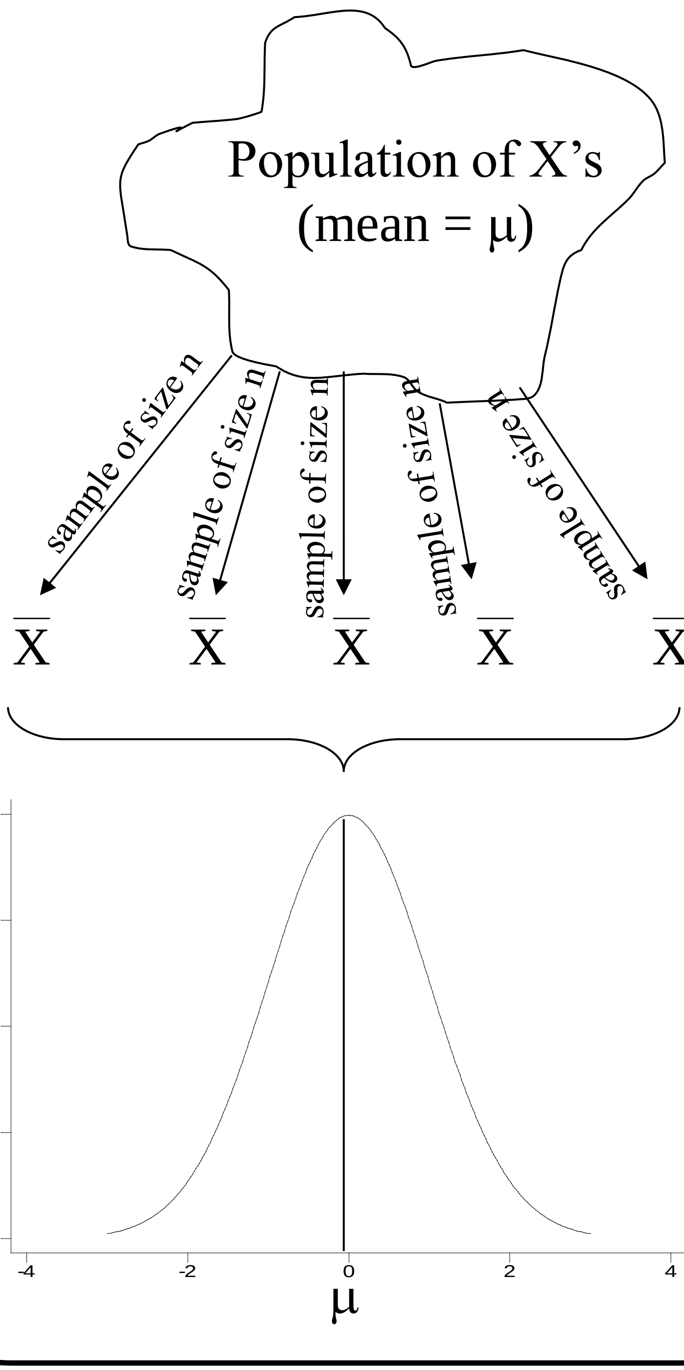
Distribution of the Sample Mean

A. When sampling from a **normally** distributed population:

1. The distribution of $\overline{X}$ is **normal**.
2. $\overline{X}$ is a random variable.
3. Mean of $\overline{X}$ is $\mu_{\overline{X}}$ which equals μ, the mean of the population.
4. Variance of $\overline{X}$ is $\sigma^2_{\overline{X}}$ which equals, σ^2/n the variance of the population divided by the sample size.
5. $\overline{X} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$.

B. When the population is **non-normal** but the sample size is large, the **Central Limit Theorem** applies.

Distribution of the Sample Mean



Central Limit Theorem

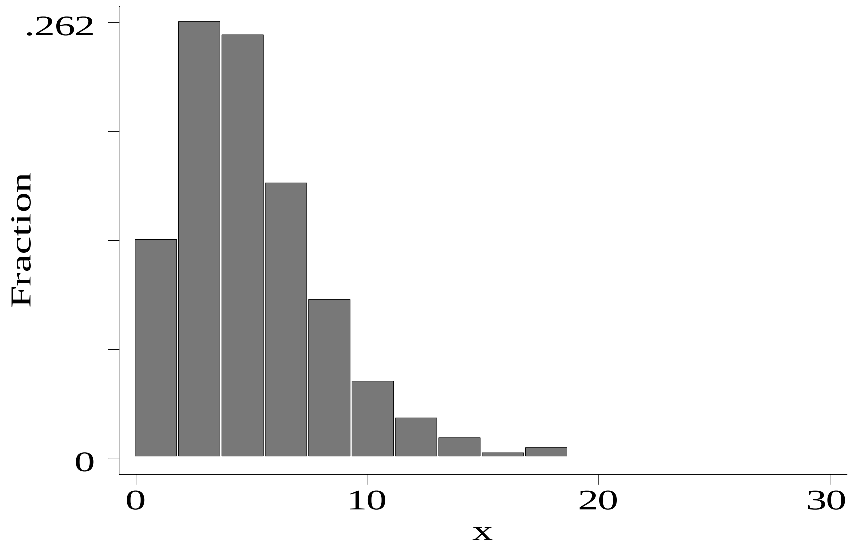
Given a population with any non-normally distributed variables with a mean μ and a variance σ^2, then for large enough sample sizes, the distribution of the sample mean, $\overline{X}$, will be **approximately normal** with means μ and variance σ^2/n.

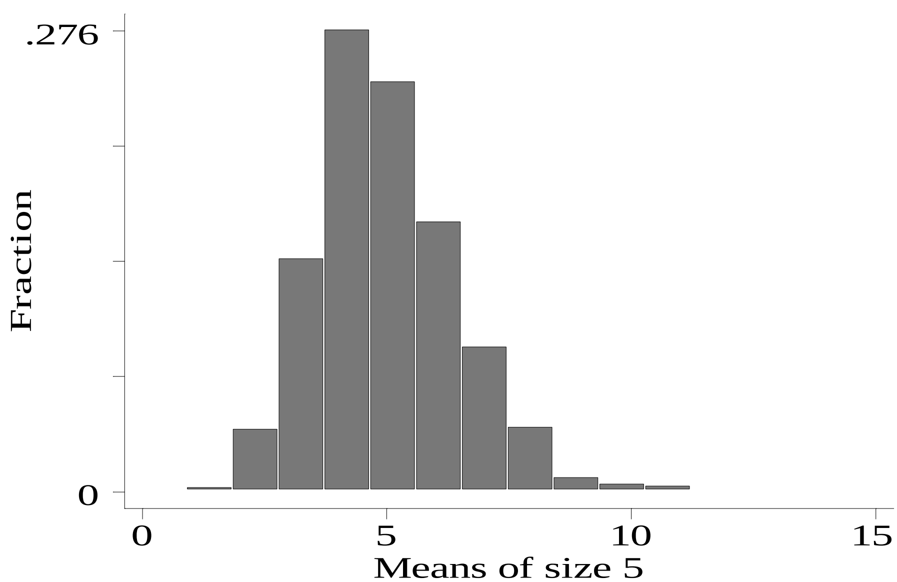
$$\text{n large} \rightarrow \overline{\mathrm{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

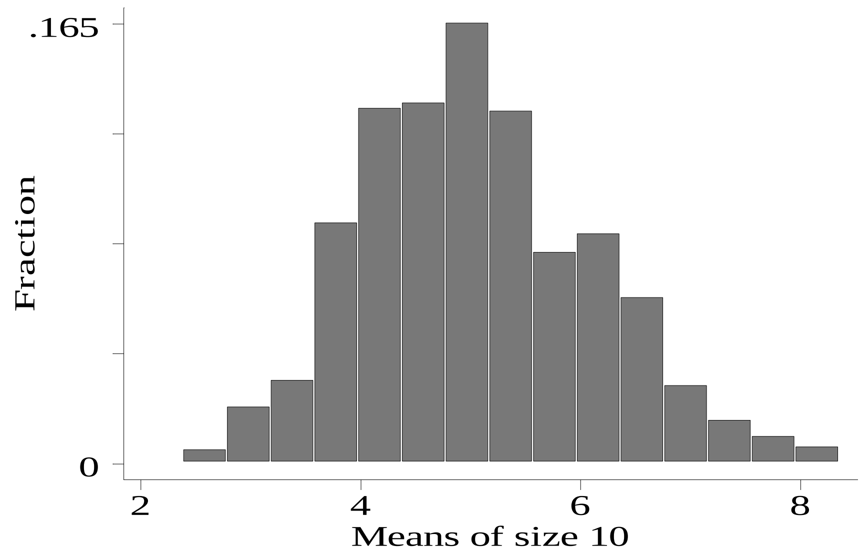
- In general, this applies for $n \geq 30$.
- As n increases, the normal approximation improves.

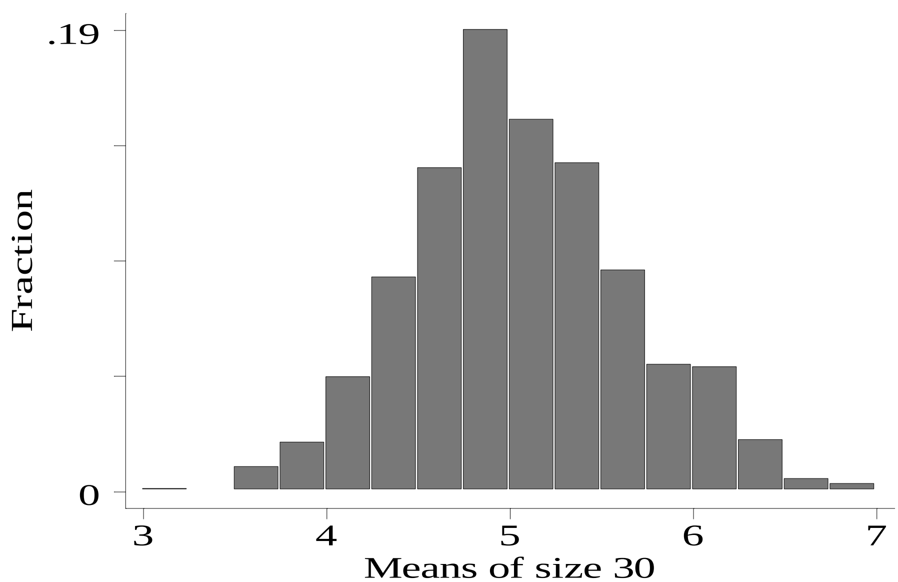
Central Limit Theorem - Illustration

Population






.165
0
Fraction
2
4
6
8
Means of size 10


.19
0
Fraction
3
4
5
6
7
Means of size 30

Distribution of Sample Mean

In applications we can address:

What is the probability of obtaining a sample with mean larger (smaller) than T (some constant) when sampling from a population with mean μ and variance σ^2?

Transform to Standard Normal

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

random variable = $\overline{X}$

distribution of sample mean ≈ Normal

expected value of sample mean = μ

standard deviation of sample mean = $\frac{\sigma}{\sqrt{n}}$

Distribution of the Sample Mean

EXAMPLE:

Suppose that for Seattle sixth grade students the mean number of missed school days is 5.4 days with a standard deviation of 2.8 days. What is the probability that a random sample of size 49 (say Ridgecrest's 6th graders) will have a mean number of missed days greater than 6 days?

<u>Random Variable</u>

<u>Distribution</u>

<u>Parameters</u>

<u>Question</u>

Find the probability that a random sample of size 49 from this population will have a mean greater than 6 days.

μ = 5.4 days

σ = 2.8 days

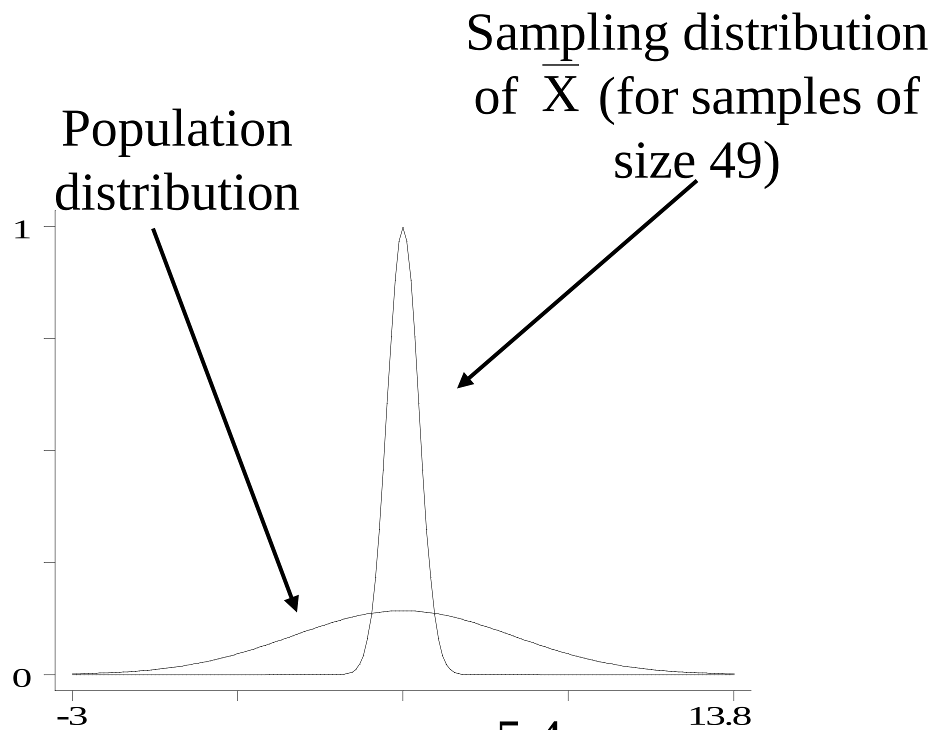
n = 49

$$\sigma_{\overline{X}} = \sigma / \sqrt{n} = 2.8 / \sqrt{49} = 0.4$$

$$\mu_{\overline{X}} = 5.4$$

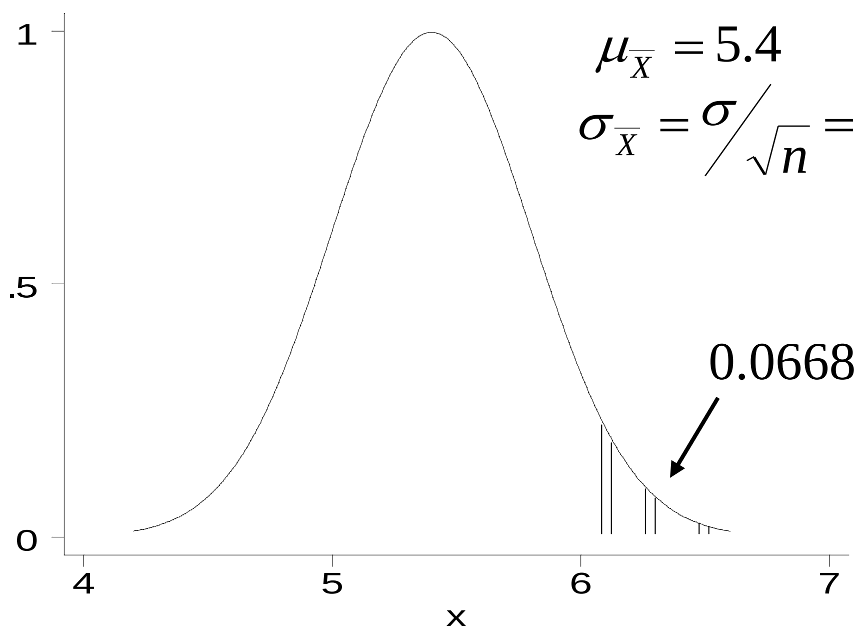
$$P\left(\overline{X} > 6\right) = P\left(\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} > \frac{6 - 5.4}{0.4}\right)$$

$$= P\left(Z > 1.5\right) = 0.0668$$



$$\mu = \mu_{\overline{X}} = 5.4$$

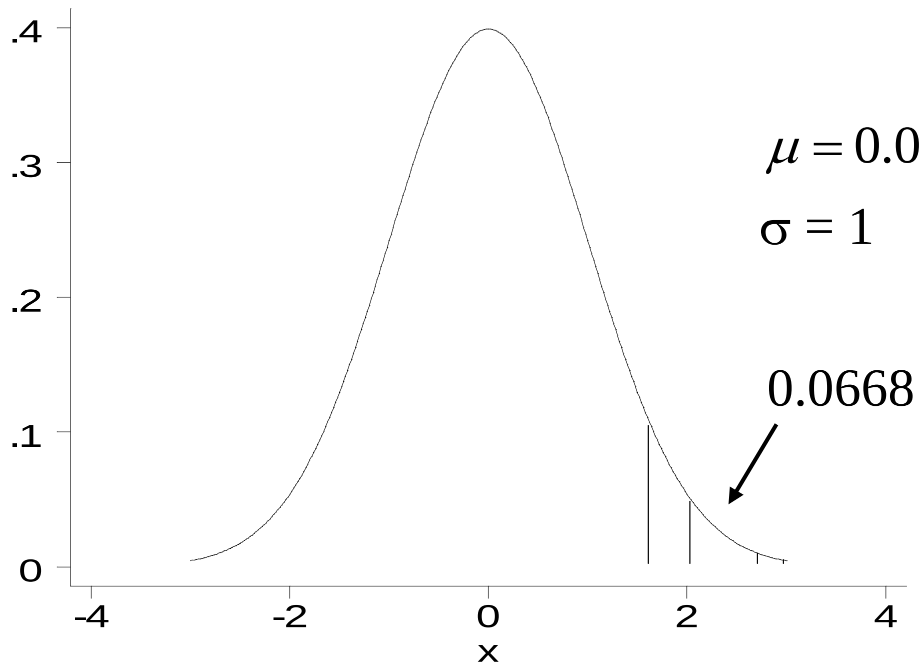
Let’s look at the sampling distribution more closely ...



$$\mu_{\overline{X}} = 5.4$$

$$\sigma_{\overline{X}} = \sigma / \sqrt{n} = 0.4$$

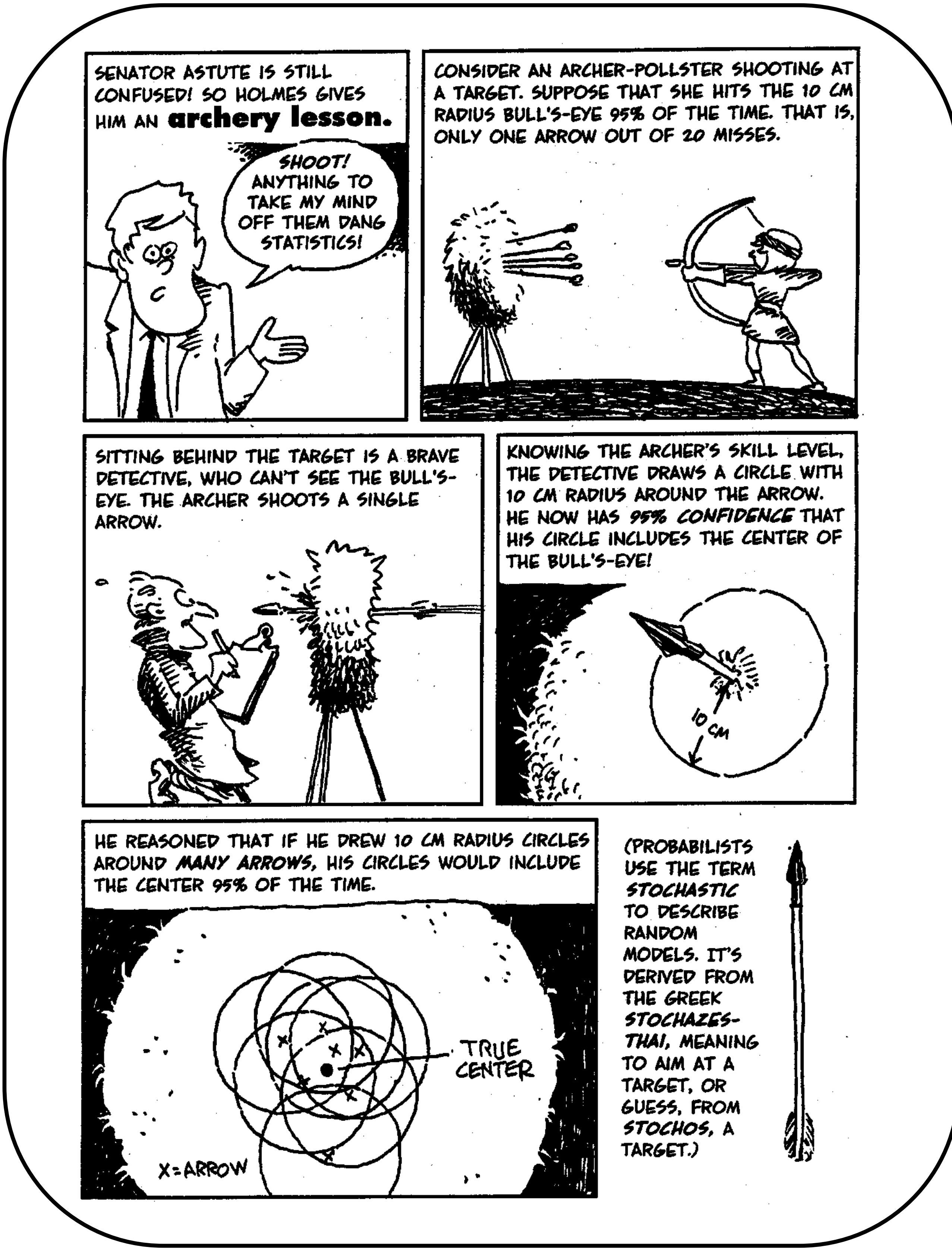
In terms of the standard normal ...



What is the probability that a random sample (size 49) from this population has a mean between 4 and 6 days? Check that ….

$$
\begin{aligned}
P(4 \le \overline{X} \le 6) &= P(-3.5 \le Z \le 1.5) \\
&= P(Z \le 1.5) - P(Z \le -3.5) \\
&= .933
\end{aligned}
$$

Confidence Intervals



SENATOR ASTUTE IS STILL CONFUSED! SO HOLMES GIVES HIM AN archery lesson.
SHOOT! ANYTHING TO TAKE MY MIND OFF THEM DANG STATISTICS!
CONSIDER AN ARCHER-POLLSTER SHOOTING AT A TARGET. SUPPOSE THAT SHE HITS THE 10 CM RADIUS BULL'S-EYE 95% OF THE TIME. THAT IS, ONLY ONE ARROW OUT OF 20 MISSES.
SITTING BEHIND THE TARGET IS A BRAVE DETECTIVE, WHO CAN'T SEE THE BULL'S-EYE. THE ARCHER SHOOTS A SINGLE ARROW.
KNOWING THE ARCHER'S SKILL LEVEL, THE DETECTIVE DRAWS A CIRCLE WITH 10 CM RADIUS AROUND THE ARROW. HE NOW HAS 95% CONFIDENCE THAT HIS CIRCLE INCLUDES THE CENTER OF THE BULL'S-EYE!
10 CM
HE REASONED THAT IF HE DREW 10 CM RADIUS CIRCLES AROUND MANY ARROWS, HIS CIRCLES WOULD INCLUDE THE CENTER 95% OF THE TIME.
TRUE CENTER
X=ARROW
(PROBABILISTS USE THE TERM STOCHASTIC TO DESCRIBE RANDOM MODELS. IT'S DERIVED FROM THE GREEK STOCHAZESTHAI, MEANING TO AIM AT A TARGET, OR GUESS, FROM STOCHOS, A TARGET.)

Confidence Intervals

Q: When we do not know the population parameter, how can we use the sample to estimate the population mean, and use our knowledge of probability to give a range of values consistent with the data?

Parameter: μ

Estimate: $\overline{X}$

Given a normal population, or large sample size, we can state:

$$P\left[-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le +1.96\right] = 0.95$$

Confidence Intervals

$$P\left[-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le +1.96\right] = 0.95$$

We can do some rearranging:

$$P\left[-1.96\sigma/\sqrt{n} \le \overline{X}-\mu \le +1.96\sigma/\sqrt{n}\right] = 0.95$$

$$P\left[-\overline{X}-1.96\sigma/\sqrt{n} \le -\mu \le -\overline{X}+1.96\sigma/\sqrt{n}\right] = 0.95$$

$$P\left[\overline{X}-1.96\sigma/\sqrt{n} \le \mu \le \overline{X}+1.96\sigma/\sqrt{n}\right] = 0.95$$

The interval

$$\left(\overline{X}-1.96\sigma/\sqrt{n}\,,\ \overline{X}+1.96\sigma/\sqrt{n}\right)$$

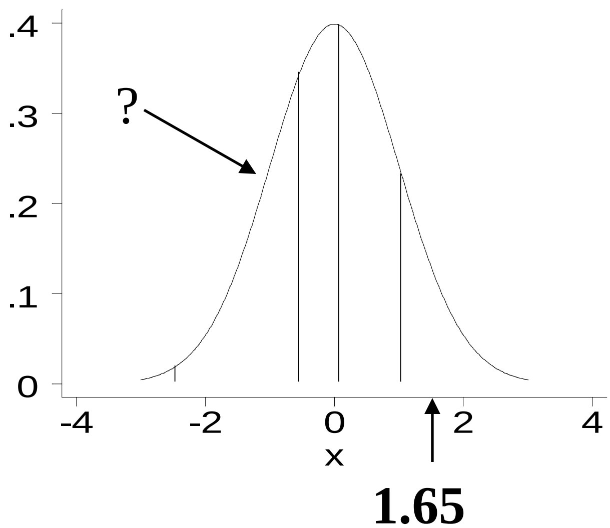
is called a **95% confidence interval for** μ.

Normal Quantiles

Go back to Rosner, table 3 ….

Notice that we can use the table two ways:

(1) Given a particular x value (the quantile) we can look up the probability:

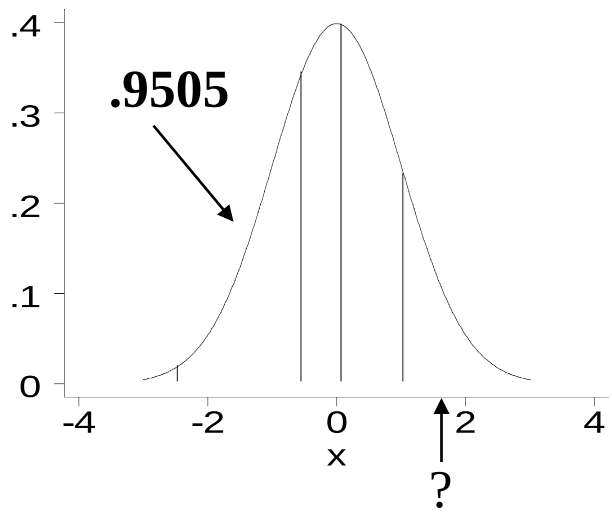


$P(Z \leq \mathbf{1.65}) = ?$

Rosner table 3 …

x	A	B	C	D
:	:	:	:	:
1.65	.9505	.0495	.4505	.9011
:	:	:	:	:

(2) Given a particular probability, we can look up the quantile:



P(Z $\leq$?) = **.9505**

x	A	B	C	D
:	:	:	:	:
1.65	**.9505**	.0495	.4505	.9011
:	:	:	:	:

$Q_Z^{(p)}$ is the value of x such that P(Z $\leq$ x) = p

Verify: $Q_Z^{(.95)} = 1.65$

$Q_Z^{(.975)} = 1.96$

$$P(Q_Z^{(.05)} \leq Z \leq Q_Z^{(.95)}) = .90 \quad \Rightarrow \quad Q_Z^{(.05)} = -1.65, \;\; Q_Z^{(.95)} = 1.65$$

Notice that $Q_Z^{(p)} = -Q_Z^{(1\text{-}p)}$

Confidence Intervals
σ known

When σ is known we can construct a confidence interval for the population mean, μ, for any given confidence level, (1 - α). Instead of using 1.96 (as with 95% CI's) we simply use a different constant that yields the right probability.

So if we desire a (1 - α) confidence interval we can derive it based on the statement

$$P\left[Q_Z^{\left(\frac{\alpha}{2}\right)} < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < Q_Z^{\left(1-\frac{\alpha}{2}\right)} \right] = 1 - \alpha$$

That is, we find constants $Q_Z^{\left(\frac{\alpha}{2}\right)}$ and $Q_Z^{\left(1-\frac{\alpha}{2}\right)}$ that have exactly (1 - α) probability between them.

A (1 - α) Confidence Interval for the Population Mean

$$\left(\overline{X} + Q_Z^{\left(\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + Q_Z^{\left(1-\frac{\alpha}{2}\right)} \times \frac{\sigma}{\sqrt{n}} \right)$$

Confidence Intervals

σ known - EXAMPLE

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of 30 second time mothers yield a mean pregnancy length of 279.5 days. Construct a 90% confidence interval for the mean length of second pregnancies based on this sample.

Confidence Intervals
σ unknown

To get a CI for μ using the methods outlined above, we need $\overline{X}$ and σ^2. But usually, **σ is unknown** - we only have $\overline{X}$ and s^2. It turns out that even though

$$\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}$$

is normally distributed,

$$\frac{(\overline{X}-\mu)}{s/\sqrt{n}}$$

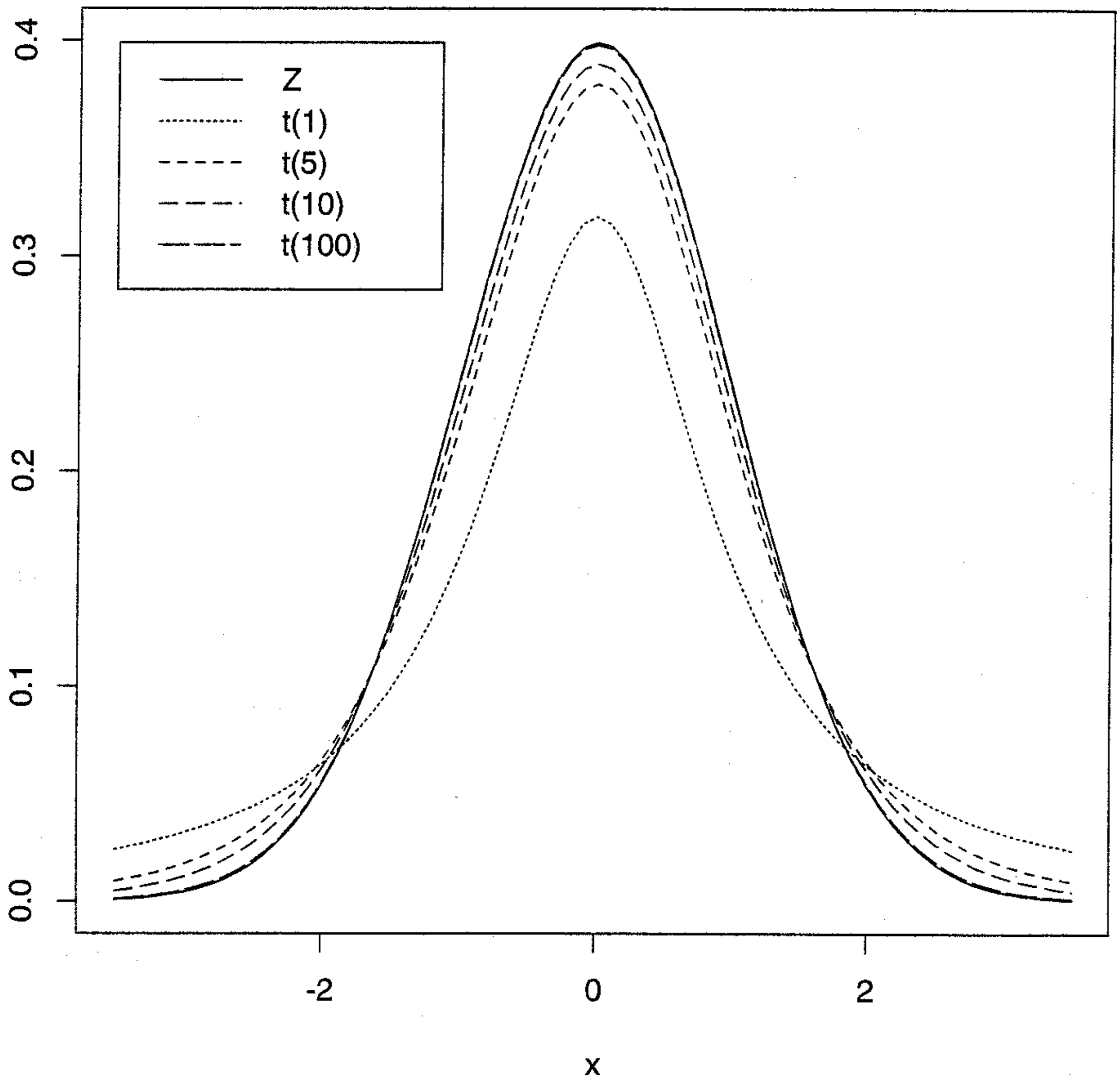
is not (quite)!

W.S. Gosset worked for Guinness Brewing in Dublin, IR. He was forced to publish under the pseudonym "Student". In 1908 he derived the distribution of

$$\frac{(\overline{X}-\mu)}{s/\sqrt{n}}$$

which is now known as Student's **t-distribution**.

Normal and t distributions



Confidence Intervals
σ^2 unknown
t Distribution

When σ is unknown we replace it with the estimate, s, and use the t-distribution. The statistic

$$\frac{\overline{X} - \mu}{s / \sqrt{n}}$$

has a t-distribution with n-1 *degrees of freedom*.

We can use this distribution to obtain a confidence interval for μ even when σ is not known.

See Rosner, table 5 or `display tprob(df,t)`

A (1-α) **Confidence Interval for the Population Mean when σ is unknown**

$$\left(\overline{X} + Q_{t(n-1)}^{\left(\frac{\alpha}{2}\right)} \times s / \sqrt{n},\ \overline{X} + Q_{t(n-1)}^{\left(1-\frac{\alpha}{2}\right)} \times s / \sqrt{n} \right)$$

Confidence Intervals - σ^2 unknown
t Distribution - EXAMPLE

Given our 30 moms with a mean gestation of 279.5 days and a variance of 28.3 days^2, we can now compute a 95% confidence interval for the mean length of pregnancies for second time mothers:

Confidence Intervals - sample variance

Q: Can we derive a confidence interval for the sample variance?

A: Yes. We'll need the **Chi-square distribution**

Definition: The sum of squared independent standard normal random is a random variable with a **Chi-square** distribution with n degrees of freedom.

Let Z_i be standard normals, N(0,1). Let

$$X = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2$$

X has a $\chi^2(n)$ **distribution**

Chi-square Distribution

Properties of χ^2 (n): Let $X \sim \chi^2(n)$.

1. $X \geq 0$
2. $E[X] = n$
3. $V[X] = 2n$
4. **n**, the parameter of the distribution is called *the degrees of freedom*.

Chi-square Distribution
Sample Variance

The Chi-square distribution describes the distribution of the **sample variance**. Recall

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

and

$$(n-1)\frac{s^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

Now the right side almost looks like

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

which would be $\chi^2(n)$.

Since μ is estimated by $\overline{X}$ one degree of freedom is lost leading to …

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2 \quad \text{with n-1 degrees of freedom}$$

Chi-square Distribution
Confidence Interval for σ²

We can use the Chi-square distribution to obtain a (1 - α) confidence interval for the **population variance**.

$$P\left[Q^{\left(\frac{\alpha}{2}\right)}_{\chi^2(n-1)} < (n-1)\frac{s^2}{\sigma^2} < Q^{\left(1-\frac{\alpha}{2}\right)}_{\chi^2(n-1)} \right] = 1-\alpha$$

Now, inverting this statement yields:

$$P\left[s^2 \times (n-1)/Q^{\left(1-\frac{\alpha}{2}\right)}_{\chi^2(n-1)} < \sigma^2 < s^2 \times (n-1)/Q^{\left(\frac{\alpha}{2}\right)}_{\chi^2(n-1)} \right] = 1-\alpha$$

Therefore,

A (1 - α) **Confidence Interval for the Population Variance**

$$\left(s^2 \times (n-1)/Q^{\left(1-\frac{\alpha}{2}\right)}_{\chi^2(n-1)},\ s^2 \times (n-1)/Q^{\frac{\alpha}{2}}_{\chi^2(n-1)} \right)$$

Chi-square Distribution
Confidence Interval for σ^2 - EXAMPLE

Suppose for the second time mothers were not happy using the standard deviation of 6 days since it was based on the population of all mothers regardless of parity. The sample variance was 28.3 days^2. What is a 95% confidence interval for the variance of the length of second pregnancies?

Summary

- General (1 - α) Confidence Intervals.
- CI for μ, σ assumed known $\rightarrow$ Z.
- CI for μ, σ unknown $\rightarrow$ T.
- CI for $\sigma^2 \rightarrow \chi^2$

- $\uparrow$confidence $\rightarrow$ wider interval
- $\uparrow$sample size $\rightarrow$ narrower interval