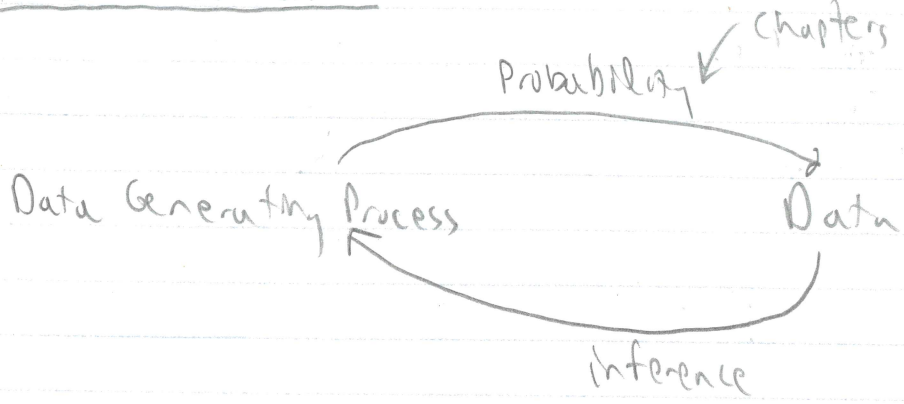
Chapter 6: Estimation



- Probability: $X \sim N(80, 12)$
  ↑ assumed know
  What is $Pr(X > 90)$

- Inference: Observe $X_1 = 81, 78, 77, 89, \ldots$
  ↑ Assume $X_i \sim N(\mu, \sigma^2)$ what are μ and σ^2?

- Estimation: Guess parameter values from data
  - Point estimation: A single number is your best guess
    ↑ E.g. estimate μ with $\bar{x}$
  - Interval estimation: Get a range of likely values of a parameter
    ↑ confidence intervals

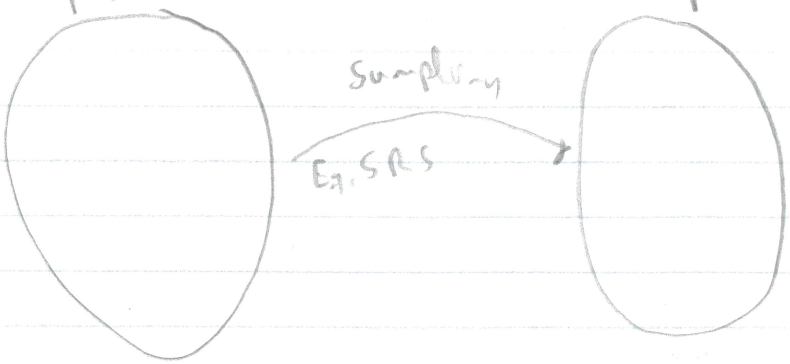
- Hypothesis Testing: How sure are we a parameter is different from some value?
  ↑ e.g. non-zero?

group we are interested in

group we have data about

Reference / target / study population

Sample



parameter: summary of pop.
μ, σ^2, p

statistic: summary of sample
$\bar{x}$, s^2, $\hat{p}$

- Simplest way to get a sample is by a <u>simple random sample</u>

  [each unit has an equal chance of being in sample

- Random selection (via SRS) is distinct from

  <u>Random Assignment</u>: Randomly assign units to different groups (eg. treatment vs control)

- Random selection: results generalizable to target population (b/c sample similar to pop in terms of demographic variables, etc.)

- Random Assignment: Allows for claims of causality because all possible confounders equal in the groups.

- Randomized Clinical Trial (RCT): Random assignment of treatment to compare them

- No causal claims without random assignment

E.g. tobramycin and gentamicin are antibiotics, tobramycin is more aggressive and has more side effects. Early studies were not randomized and showed tobramycin performed worse. Why?

↳ Doctors gave sicker patients tobramycin

↳ randomization guarantees equal # sicker and less sick in each group (on average).

best
- Double blind – neither doctor nor patient know treatment

↳ guards against placebo effect

ok
- Single blind – doctor knows

worst
- unblinded – both know

- Sampling in R

- Estimate mean:

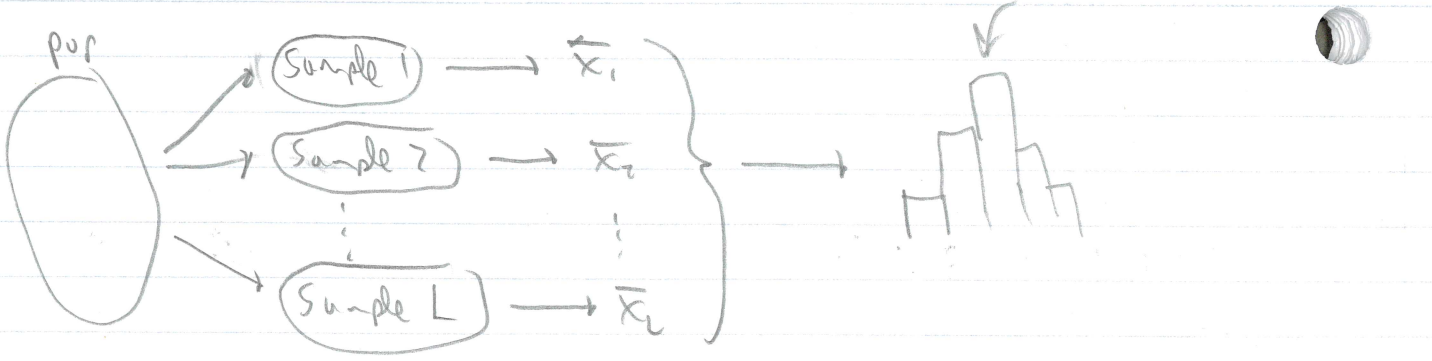
Suppose $E[X_i] = \mu$

↑ Not necessarily normal

Estimate μ with $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\hat{\mu}$: estimate of μ; $\bar{x}$: sample mean

- Sampling distribution: Distribution of statistic across many (hypothetical) samples.



↑ used to describe properties of statistics

- Properties of $\bar{X}$

Let $X_1, X_2, \ldots, X_n$ be a random sample with $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2$

Then

1.) $E(\bar{X}) = \mu$ (unbised)

2.) $\text{Var}(\bar{X}) = \sigma^2/n$

↳ more precise for larger n (consistent)

3.) $\bar{X} \approx N(\mu, \sigma^2/n)$ for large n even if X_i are not also normal

(central limit theorem)

- Standard error: standard deviation of a statistic

$SE(\bar{X}) = \sigma/\sqrt{n}$ ↑ Not individual observations

- Estimated Standard Error (Standard Error): estimated standard deviation of a statistic

↳ Book just means estimated SE when they say SE, typically

$SE(\bar{X}) = s/\sqrt{n}$

- ↳ typically don't know σ^2, so estimate it with s^2

Statistic	X_i	$\bar{X}$
Mean	μ	μ
Variance	σ^2	σ^2/n
Estimated variance	s^2	s^2/n
Distribution	Unknown	$N(\mu, \sigma^2/n)$ (for large n)

- Exercise: Mean birthweight is 112 oz with standard deviation 20.6. What is the probability the mean of 10 birthweights will be between 98 and 126?

$$\bar{X} \sim N\left(112, \left(20.6/\sqrt{10}\right)^2\right) = N(112, 6.514^2)$$

$$= \text{pnorm}(126, 112, 6.514) - \text{pnorm}(98, 112, 6.514)$$

$$= 0.9684$$

Interval Estimation

- $\bar{X}$ is not exactly equal to μ
- Want range of likely values of μ
- Know $\bar{X} \sim N(\mu, \sigma^2/n)$ for large n

$$\Rightarrow Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95$$

$$\Rightarrow Pr\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$\Rightarrow Pr\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$\Rightarrow \bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$ captures μ w.p. 0.95

↑ 95% CI

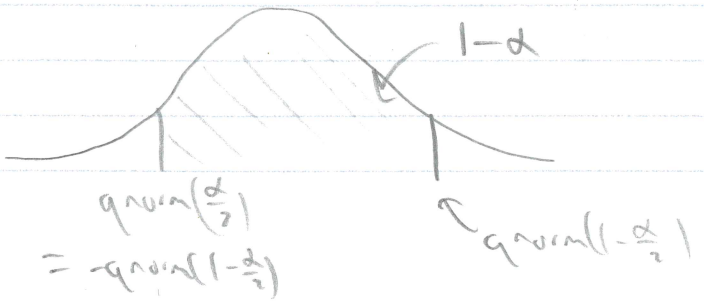
- Estimate ± multiplier × standard error

↑ common CI format

- More generally,

$$\boxed{\bar{X} \pm qnorm\left(1-\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}}$$

$(1-\alpha)100\%$ CI for μ



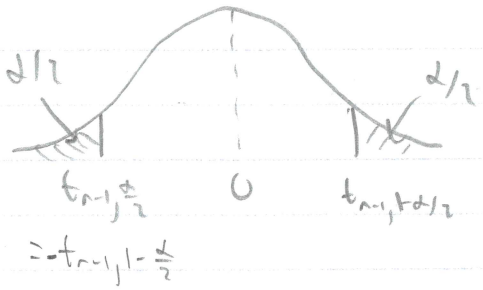
R Interpretation of CI

- The above only works when σ^2 is known
  - ↕ σ^2 is never known.
- $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ (Not $N(0,1)$!)
  - ↕ t-distribution with $n-1$ degrees of freedom
- Only an exact result when $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$
  - ↕ But t distribution tends to work better in small samples even when X_i are not normal
- For large n, $t_{n-1} \approx N(0,1)$, so CLT is OK.
- Bell shaped, centered at 0
- df ↘ ⇒ extreme values more likely,
  df ↗ ⇒ extreme values less likely
- Use t b/c of added variability from using s^2 instead of σ^2
- R code for t-distribution

- Rosner uses notation $t_{df,p}$ for the p quantile of a t_{df} distribution.

$$t_{df,p} = qt(p, df)$$

- $$Pr\left(-t_{n-1,1-\frac{\alpha}{2}} \le \frac{\bar{X}-\mu}{s/\sqrt{n}} \le t_{n-1,1-\alpha/2}\right) = 1-\alpha$$



$$\Rightarrow Pr\left(-t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \le \bar{X}-\mu \le t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

$$\Rightarrow Pr\left(\bar{X} - t_{n-1,1-\frac{\alpha}{2}} \le \mu \le \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Rightarrow \boxed{\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}}$$

$\uparrow$ $(1-\alpha)100\%$ CI for μ

- Exercise: $n = 10$, $\bar{x} = 116.90$, $s = 21.70$

Calculate 90%, 95%, 99% CI's

Solution: $qt(0.95, 9) = 1.833$
$qt(0.975, 9) = 2.262$
$qt(0.995, 9) = 3.25$

$116.96 \pm 1.833 \cdot 21.70/\sqrt{10}$

$116.96 \pm 2.262 \cdot 21.70/\sqrt{10}$

$116.96 \pm 3.25 \cdot 21.70/\sqrt{10}$

- Note: CI level ↑ (so α ↓) $\Rightarrow$ larger intervals

  n ↑ $\Rightarrow$ smaller intervals

  s^2 ↑ $\Rightarrow$ larger intervals

- **Bone Density Case Study**

- Estimate Variance:

  Estimate σ^2 with $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Why $n-1$?

  $\bar{x} = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n}(x_i - a)^2$

  $\Rightarrow \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{too small}} \leq \overbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}^{\text{ideal}}$

  $\therefore$ Dividing by $n-1$ makes it bigger.

- $\frac{S^2}{\sigma^2/(n-1)} \sim \chi^2_{n-1}$

↑ Chi-squared distribution with $n-1$ degrees of freedom

- Properties of Chi-squared
  1.) Support ≥ 0
  2.) $E[\chi^2_{df}] = df$
  3.) $df \searrow \Rightarrow$ thicker tails (extreme events happen more often)

- Can use this to get confidence intervals for σ^2

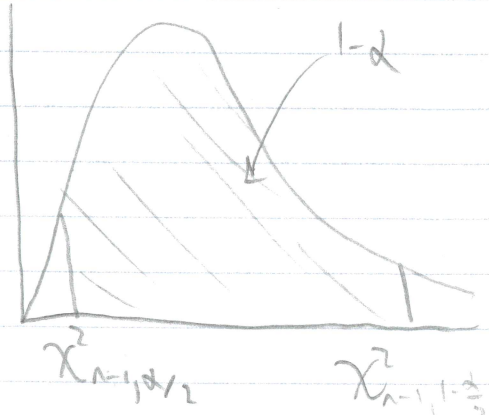
- Let $\chi^2_{df,p}$ = p^{th} quantile of a χ^2_{df} distribution

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

$\Rightarrow Pr\left[(n-1)S^2/\chi^2_{n-1,1-\alpha/2} \leq \sigma^2 \leq (n-1)S^2/\chi^2_{n-1,\alpha/2}\right] = 1-\alpha$

$$\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}$$

$100(1-\alpha)\%$ CI



- Notes: - Less often constructed than CI for μ
  - Very sensitive to violations of normality
    ↳ CLT does not save you

- $X_1, X_2, \ldots, X_n$ be independent Bernoulli trials s.t.

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

Then $X = \sum_{i=1}^{n} X_i$ = # of 1's out of n trials

$X \sim Bin(n, p)$

- Example: Estimate prevalence of malignant melanoma
  Sample $n = 5000$ individuals. $X_i = \begin{cases} 1 & \text{if has melanoma} \\ 0 & \text{if not} \end{cases}$

- Goal: Observe X, estimate p

- $\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} X$ = proportion of 1's

- $E(\hat{p}) = E\left[\frac{1}{n} X\right] = \frac{1}{n} E(X) \overset{\text{Binomial mean}}{=} \frac{1}{n} np = p \Rightarrow$ unbiased

- $Var(\hat{p}) = Var\left(\frac{1}{n} X\right) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$

↳ more precise for larger $n \Rightarrow$ consistent

- Estimated standard error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- Goal: Interval estimate of p

- For large n, $\hat{p} \sim N(p, \frac{p(1-p)}{n})$

  ↑ $n\hat{p}(1-\hat{p}) \geq 5$

- Let $z_p = \text{qnorm}(p) = p^{th}$ quantile of $N(0,1)$

- $$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1)$$

$$\Rightarrow \Pr\left(-z_{1-\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 0.95$$

Solve for p

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$ is a $(1-\alpha)100\%$ CI for p

estimate — multiplier — standard error

- Exercise: 10,000 women, 400 have breast cancer. What is a 95% CI for breast cancer incidence?

Solution: $\hat{p} = 0.04$, $n\hat{p}(1-\hat{p}) = 3847.5$,

$z_{0.975} = \text{qnorm}(0.975) = 1.96$

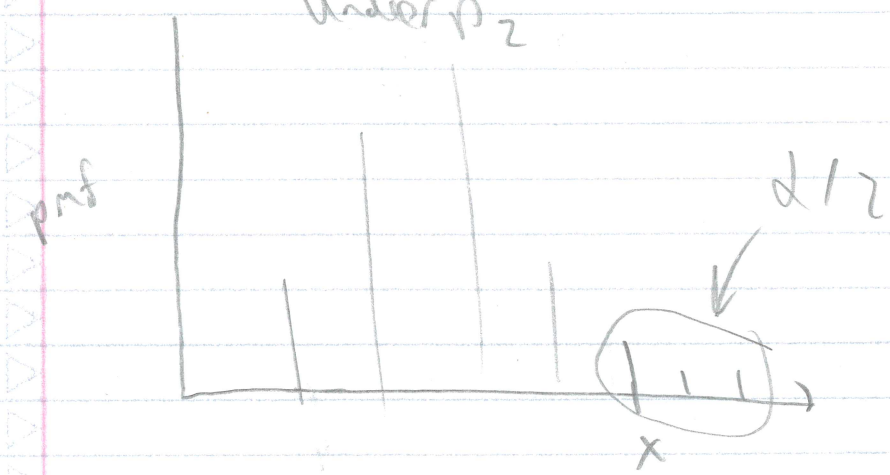
$$0.04 \pm 1.96 \cdot \sqrt{0.04 \cdot ((-0.04)/10{,}000} = (0.03616, 0.04384)$$

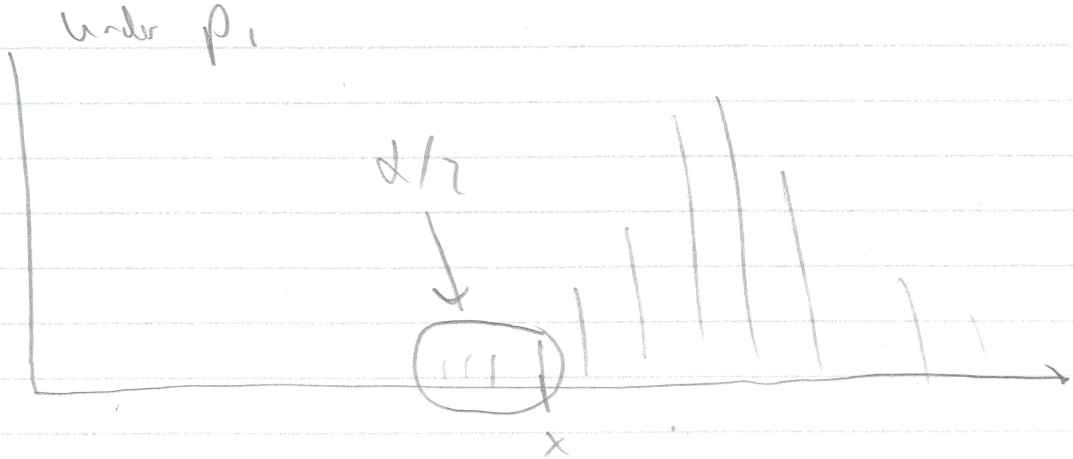
- What if n is small? Use an exact method

find p_1 and p_2 s.t.

$$\frac{\alpha}{2} = \text{pbinom}(x, \text{size}=n, \text{prob}=p_1)$$

$$\frac{\alpha}{2} = 1 - \text{pbinom}(x-1, \text{size}=n, \text{prob}=p_2)$$





Binomial R code

- Poisson Estimation

$X \sim Poi(\mu)$ where $\mu = \lambda T$

Goal: Estimate λ

λ = Incidence rate per unit time

T = time

- Ex.) Leukemia rate in Woburn, MA. 12,000 residents, 10 years. 12 children got Leukemia.

- A common unit of time in Biostatistics is a person-year
  - 1 person being followed for 1 year
  - Normalizes by # people in a study
  - Eg. 12 cases out of 20 is a lot more than 12 out of 20,000

- Ex.) Woburn study had 12000 people · 10 years = 120,000 person-years.

- $\hat{\lambda} = \frac{X}{T}$, $E(\hat{\lambda}) = \frac{E(X)}{T} = \frac{\lambda T}{T} = \lambda$ ↑ unbiased

- Woburn: $\hat{\lambda} = \frac{12 \text{ cases}}{120{,}000 \text{ person years}} = 0.0001 \frac{\text{cases}}{\text{person-year}}$

- Cancer rates low so usual to multiply by 100,000

$$\frac{0.0001 \text{ cases}}{\text{person-year}} = 10 \frac{\text{cases}}{100{,}000 \text{ person-years}}$$

- Interval same strategy as exact Binomial case

μ_1 s.t. $Pr(X \geq x \mid \mu_1) = \alpha/2$

μ_2 s.t. $Pr(X \leq x \mid \mu_2) = \alpha/2$

Interval (μ_1, μ_2) ← CDF of Poisson

- Interval for λ is $\left(\frac{\mu_1}{T}, \frac{\mu_2}{T}\right)$

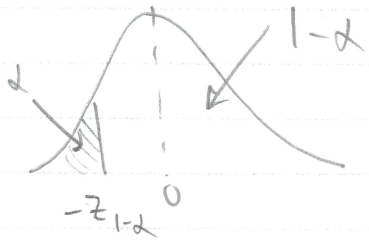
- Use poisson.test() in R

- **One-sided CI** (UK to skip)
  - only interested in lower or upper bounds
  - Eg. compare cancer treatment survival rate against baseline of 30%. Want lower bound on treatment effect to see if better than baseline

- Find p_1 s.t. $P_r(p > p_1) = 1-\alpha$ (true parameter p; p_1 random lower bound)

- Or, find p_2 s.t. $P_r(p < p_2) = 1-\alpha$
  - Eg. upper bound on incidence rate of some treatment group

- Since $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$



$$P_r\left(-z_{1-\alpha} < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}\right) = 1-\alpha$$

$$\Rightarrow P_r\left(z_{1-\alpha} > \frac{p-\hat{p}}{\sqrt{\hat{p}(1-\hat{p})/n}}\right) = 1-\alpha$$

$$\Rightarrow P_r\left(\hat{p} + z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n} > p\right) = 1-\alpha$$

$$\Rightarrow p_2 = \hat{p} + z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n}$$

Similarly $p_1 = \hat{p} - z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n}$

Ex.) 40 out of 100 patients survive a new cancer treatment. What is lower bound on $Pr(\text{survive})$? (Upper 1-sided 95% CI)

$$\hat{p} = \frac{40}{100} = 0.4$$

$$z_{1-\alpha} = qnorm(0.95) = 1.645$$

$\uparrow_{\alpha = 0.05}$

$$\hat{p} - z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.4 - 1.645\sqrt{\frac{0.4 \cdot 0.6}{100}}$$

$$= 0.319$$