Intro. to Probability

Probability: Measure of uncertainty or likelihood associated with the occurrences or outcomes of events

$$P(z_i) \approx \frac{m}{N}$$

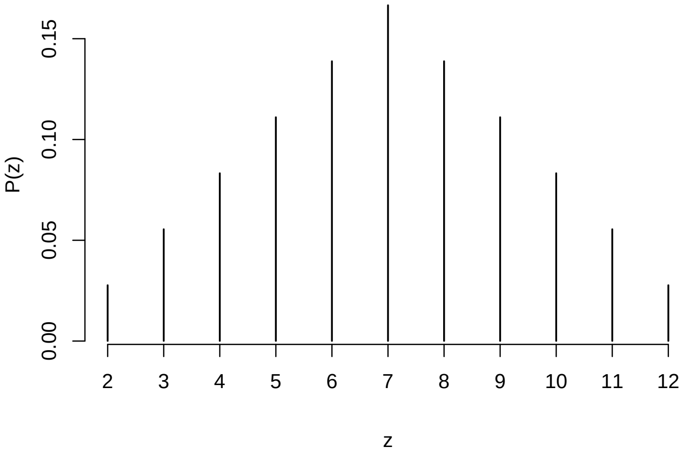
- N total number of all possible outcomes (**sample space**)
- m number of times a given outcome (**event**, z_i) is observed

Probability: Measure of uncertainty or likelihood associated with the occurrences or outcomes of events

$$P(z_i) \approx \frac{m}{N}$$

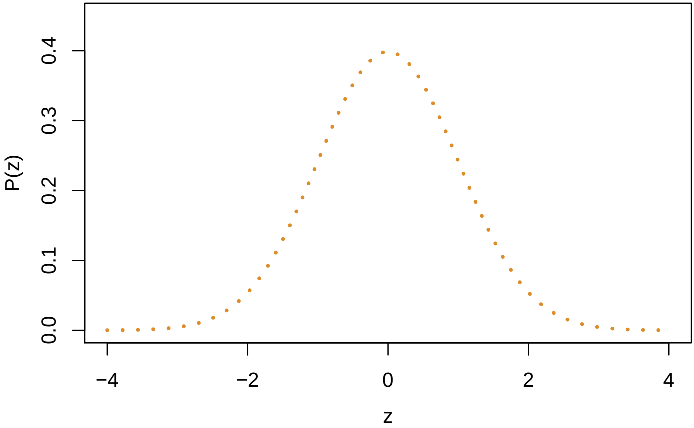
- N total number of all possible outcomes (**sample space**)
- m number of times a given outcome (**event**, z_i) is observed
- $P(2) = 1/36$
- $P(3) = 2/36$
- $P(12) = 1/36$

For discrete variables



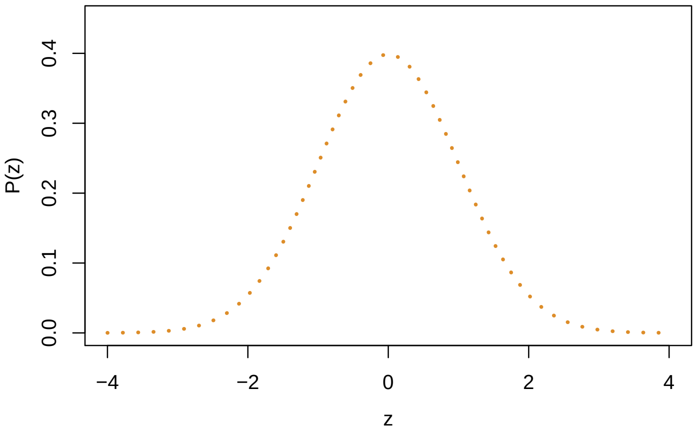
$$\sum_i P(z = z_i) = 1$$

For continuous variables



$$\int_{z_{min}}^{z_{max}} p(z)dz = 1$$

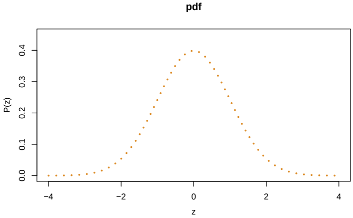
For continuous variables

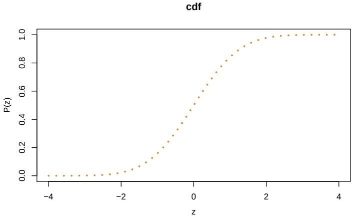


Continuous variables can be measured with arbitrary precision, thus compute probabilities over an interval

$$P(z_1 \leq z \leq z_2) = \int_{z_1}^{z_2} p(z)dz$$

- **Probability density function (PDF):** Function ($f(x)$) of a continuous random variable that gives the probability (area under curve) for a given interval
  - **Random variable:** any variable whose value depends on an unknown event
  - $f(x) > 0$
  - $\int f(x)dx = 1$
- **Cumulative density function:** Gives the total probability of being less than some value



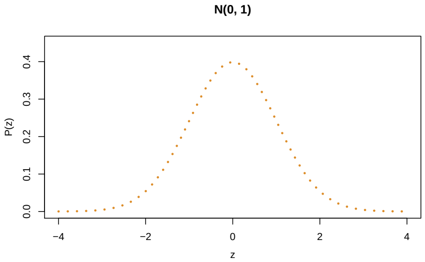


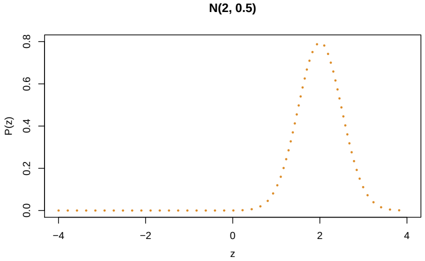
The normal PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

Parameterized by:

- Mean (μ): central location
- Variance (σ^2): spread
- $\mathrm{N}(\mu, \sigma^2)$





Statistical Inference

Statistical inference

What can we learn about the population from the sample?

Population

μ

Sample

$\bar{X}$

Statistical inference

What can we learn about the population from the sample?

Population

μ

Sample

$\bar{X}$

Goal: Estimate population mean (μ) for height from a sample of the population

Statistical inference - maximum likelihood

Goal is to estimate some unknown parameters (θ)

- Write the probability of the data in terms of θ
  - The probability density function and likelihood function are given by the same equation.
  - PDF is a function of the data when parameters are fixed
  - Likelihood function is a function of the parameters when the data is fixed
- Estimate $\hat{\theta}$, so that $\hat{\theta}$ maximizes the probability of the observed data
  - $\hat{\theta}$ is the maximum likelihood estimate (MLE)

Statistical inference - MLE of mean

Goal is to estimate the mean of a normal distribution

Recall, the normal PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

Parameterized by:

- Mean (μ): central location
- Variance (σ^2): spread
- $\mathrm{N}(\mu, \sigma^2)$

Statistical inference - MLE of mean

Goal is to estimate the mean of a normal distribution

Recall, the normal PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

and the likelihood function is

$$L(\mu, \sigma^2 | \mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Statistical inference - MLE of mean

Goal is to estimate the mean of a normal distribution

It is easier to work with the log likelihood, so the log likelihood function is

$$
\begin{aligned}
LL(\mu, \sigma^2 | \mathbf{x}) &= \sum_{i=1}^{n} log(f(x_i)) \\
&= \sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]\right) \\
&= -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned}
$$

Statistical inference - MLE of mean

Goal is to estimate the mean of a normal distribution

We can take the partial derivative of the log likelihood function with respect to μ

$$
\begin{aligned}
\frac{\partial LL}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-\frac{N}{2} log(2\pi\sigma^2)\right) + \frac{\partial}{\partial \mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)
\end{aligned}
$$

Statistical inference - MLE of mean

Goal is to estimate the mean of a normal distribution

Finally, set this to zero and solve for μ

$$\begin{aligned}
0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\
&= \sum_{i=1}^{n} x_i - N\mu \\
N\mu &= \sum_{i=1}^{n} x_i \\
\mu &= \frac{\sum_{i=1}^{n} x_i}{N}
\end{aligned}$$

Statistical inference - MLE of mean

Goal is to estimate the mean of a normal distribution

- $\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$ is the MLE for μ
- Although this is intuitive, other parameters can be obtained using a similar approach

Statistical inference

What can we learn about the population from the sample?

Population

μ

Sample

$\bar{X}$

Goal: Estimate population mean (μ) for height from a sample of the population

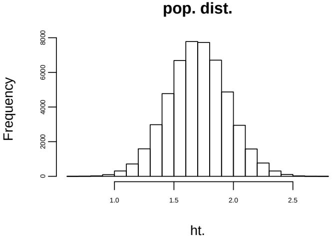
Sampling distributions

- Say we draw a random sample ($n = 18$) from the population and calculate the mean, then draw a new sample and compute the mean. Will the sample means be the same?

Sampling distributions

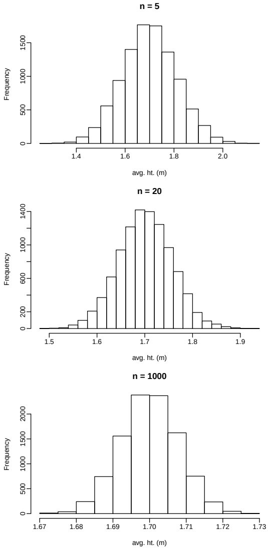
- Say we draw a random sample ($n = 18$) from the population and calculate the mean, then draw a new sample and compute the mean. Will the sample means be the same?
- The **sampling distribution** is a theoretical probability distribution of the values of some sample statistic that would occur if we were to draw all possible samples of a given size from a population.
  - Can give us insight into how common or rare it is to observe a given value for that statistic.

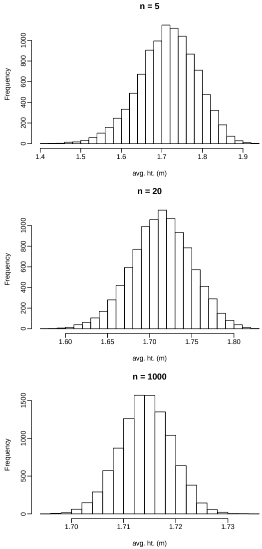
Sampling distribution of the mean

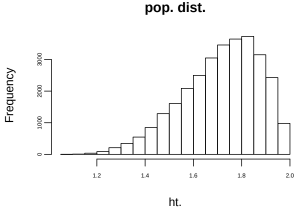


If the distribution of ht. is normal:

- $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$
- n is sample size







If the distribution of ht. is non-normal:

- For large n the distribution of the sample means will be approximately normal
- CLT states the sum of a large number of independent identically distributed random variables with a finite variance will be approximately normal

Statistical Inference - Z-test

What is the probability that our sample mean ($\bar{x} = 1.702, n = 50$) *is different from the population mean* ($\mathrm{N}(1.700, 0.0625)$)*?*

1. Create a standardized statistic ($N(0, 1)$)

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{(n)}} = \frac{0.002}{0.0353} = 0.0566$$

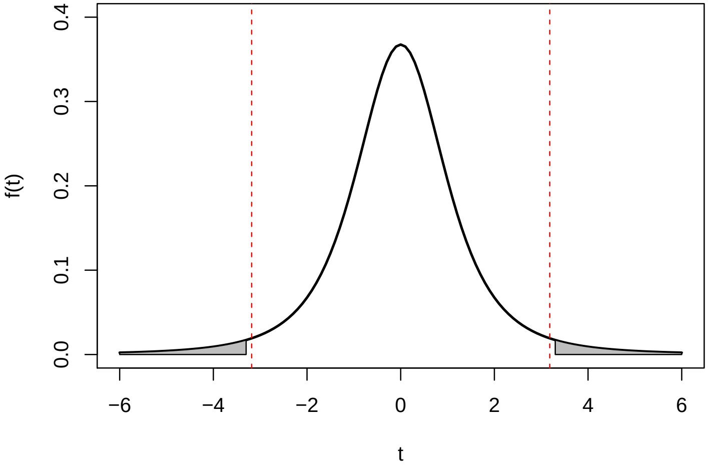
2. Compare observed stat. to null sampling distribution

$$P[x > 1.7] \int_{0.0566}^{\infty} \frac{1}{0.25\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(1.702 - 1.7)^2}{0.25^2}\right] = 0.477$$

- Since it is a two-sided test: $p = 0.477 \times 2 = 0.955$

p-values

How likely are we to observe a test statistic greater than observed value?



df = 3, $\alpha = 0.05$, $t_{obs.} = 3.275$

How precise is our estimate?

For sample mean:

$$\mathrm{SE}[\bar{x}] = \frac{s}{\sqrt{n}}$$

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

For other parameter estimates??

How precise is our estimate?

For sample mean:

$$\mathrm{SE}[\bar{x}] = \frac{s}{\sqrt{n}}$$

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

For other parameter estimates?? Resampling!

- Bootstrap: Iterative resampling
- Jackknife: Leave-one-out

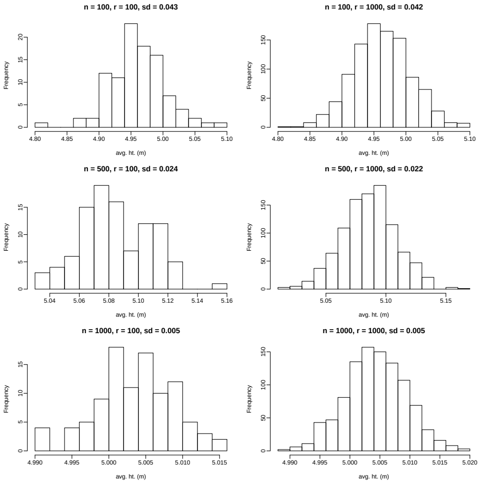
Bootstrap resampling

Given a large dataset, that adequately represents the population:

1. Randomly draw samples from the data with replacement
2. Estimate the mean (or other parameter) of this “new” dataset and record estimate
3. Repeat desired number of times
4. Calculate the standard deviation of the estimates

Bootstrap resampling

Original sample size (n) is more important than the number of iterations (r)



- Three datasets with n= 100, 500, 10000
- Mean estimated from 80% of observations
- Number of resampling iterations: r = 100, 1000

Jackknife resampling

Given a small dataset:

1. Sequentially delete one observation
2. Estimate parameter on "new" dataset
3. The standard error of the estimate is

$$\mathrm{SE}[\bar{x}] = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{x}_{(\cdot)})^2}$$

- $\bar{x}_{(\cdot)}$ is the mean of the jackknife replicates

$$\bar{x}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_i$$

Hypothesis testing

One sample t-test

- Z-test requires either a large sample size or knowledge of the sample mean and variance
  - Many cases this is unknown
- If we have prior knowledge of the mean (μ_0), then we can test to see if the sample mean ($\bar{x}$) is different from this value

One sample t-test

- Z-test requires either a large sample size or knowledge of the sample mean and variance
  - Many cases this is unknown
- If we have prior knowledge of the mean (μ_0), then we can test to see if the sample mean ($\bar{x}$) is different from this value

- **Test statistic (t-stat):**

$$t = \frac{\bar{x} - \mu_0}{\mathrm{SE}}$$

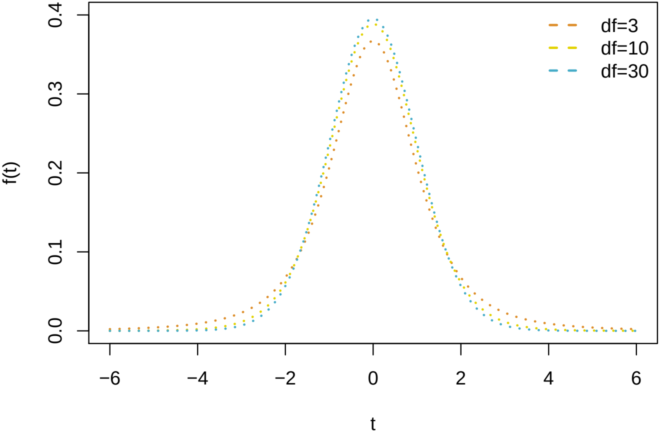
  - Standard error (SE): $\frac{s}{\sqrt{n}}$

One sample t-test

We estimate SE from the sample, need a distribution that accounts for uncertainty in estimate from sample size

One sample t-test

We estimate SE from the sample, need a distribution that accounts for uncertainty in estimate from sample size



The t-distribution is only indexed by df

- Larger sample size $\rightarrow$ s is closer to σ

Degrees of Freedom

Statistical currency: Earned by collecting independent observations, spent estimating population parameters or test statistics

- Number of pieces of new information that go into estimating some statistic
- Amount of useful information

Confidence intervals

Plausible range for parameter of interest.

- "X% chance that the interval contains the true value of the parameter"

$$\bar{x} \pm t_{(n-1),\alpha/2}\mathrm{SE}[\bar{x}]$$